NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

MEGACITY ANALYSIS: A CLUSTERING APPROACH TO CLASSIFICATION

by

Jimmy L. Housley

June 2017

Thesis Advisor:	Lyn R. Whitaker
Co-Advisor:	Jeffrey A. Appleget
Second Reader:	Robert Burks

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704–0188*

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2017	**3. REPORT TYPE AND DATES COVERED** Master’s thesis

4. TITLE AND SUBTITLE
MEGACITY ANALYSIS: A CLUSTERING APPROACH TO CLASSIFICATION

5. FUNDING NUMBERS

6. AUTHOR(S) Jimmy L. Housley

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Joint Warfare Analysis Center
4048 Higley Rd.
Dahlgren, VA 22448-5144

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

“Megacities” are characterized by large populations (at least 10 million) and interdependent infrastructure, demographic, economic, and government networks (the four pillars). To be successful in future operations, the military must expand its understanding of megacities and their networks. In particular the Joint Warfare Analysis Center (JWAC) is interested in these megacity networks and their implications for potential urban operations. We develop a methodology to group like megacities into five clusters. With 33 variables describing the four pillars, we construct a data set using over 90 data sources for 41 large urban areas. This work greatly expands previous work in both the number of cities studied and the number of variables used. We also study clustering sensitivity to missing values by generating an ensemble of 5,000 clusterings based on randomly imputed missing values. We compare these to clustering without imputation, the ensemble consensus or average clustering, and clusterings from previous studies in addition to identifying which cities are sensitive to missing values. Our work not only informs JWAC of the similarities and differences between the 41 cities studied, it provides a method to identify for which cities, more data collection is warranted, and it provides a blueprint for future work in this area.

14. SUBJECT TERMS
megacity, urban, cluster, unsupervised learning

15. NUMBER OF PAGES
153

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540–01-280-5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

MEGACITY ANALYSIS: A CLUSTERING APPROACH TO CLASSIFICATION

Jimmy L. Housley
Captain, United States Marine Corps
B.S., United States Naval Academy, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2017

Approved by: Lyn R. Whitaker
Thesis Advisor

Jeffrey A. Appleget
Co-Advisor

Robert Burks
Second Reader

Patricia A. Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

"Megacities" are characterized by large populations (at least 10 million) and interdependent infrastructure, demographic, economic, and government networks (the four pillars). To be successful in future operations, the military must expand its understanding of megacities and their networks. In particular the Joint Warfare Analysis Center (JWAC) is interested in these megacity networks and their implications for potential urban operations. We develop a methodology to group like megacities into five clusters. With 33 variables describing the four pillars, we construct a data set using over 90 data sources for 41 large urban areas. This work greatly expands previous work in both the number of cities studied and the number of variables used. We also study clustering sensitivity to missing values by generating an ensemble of 5,000 clusterings based on randomly imputed missing values. We compare these to clustering without imputation, the ensemble consensus or average clustering, and clusterings from previous studies in addition to identifying which cities are sensitive to missing values. Our work not only informs JWAC of the similarities and differences between the 41 cities studied, it provides a method to identify for which cities, more data collection is warranted, and it provides a blueprint for future work in this area.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AFRICOM	Africa Command
AOI	arc of instability
AOR	area of responsibility
BBC	British Broadcasting Corporation
CENTCOM	Central Command
CIA	Central Intelligence Agency
CLARA	clustering large applications
COCOM	combatant command
COIN	counter-insurgency
DHS	Department of Homeland Security
DOD	Department of Defense
EUCOM	European Command
FCC	Federal Communications Commission
FOE	future operating environment
GDP	gross domestic product
GMP	gross metropolitan product
GRDP	gross regional domestic product
HA/DR	humanitarian assistance/disaster relief
HET	heavy equipment transport
ICS-CERT	Industrial Control Systems Cyber Emergency Response Team
IMO	International Maritime Organization
IO	information operations
ISIS	Islamic State of Iraq and Syria
JWAC	Joint Warfare Analysis Center
K-NN	k-nearest neighbor
MCSEF	Marine Corps Security Environment Forecast
MICE	Multiple Imputation by Chained Equations
MNAR	missing not at random
MSA	metropolitan statistical area
MW	megawatts

NORTHCOM	Northern Command
PACOM	Pacific Command
PAM	partitioning around medoids
PMESII-PT	political, military, economic, social, information, infrastructure, physical environment, and time
SOUTHCOM	Southern Command
TEU	twenty-foot equivalent unit
UN	United Nations

EXECUTIVE SUMMARY

As the U.S. military prepares for the future, senior leaders and analysts alike expect that urban environments will play an increasing role in the operations we conduct, as outlined in the U.S. Army's Strategic Studies Group report on six megacity case studies (United States Army, 2014). People are migrating to urban areas and the littorals at an increasing rate, which makes understanding the structure and unique challenges of working in these densely populated regions important. These metropolitan areas are characterized by large populations (at least 10 million) and interdependent infrastructure, demographic, economic, and government networks (the four pillars). As a result, the Joint Warfare Analysis Center (JWAC) is interested in identifying how these interdependent networks influence a megacity due to the implications of potential kinetic or non-kinetic urban operations.

We develop and implement a methodology to classify megacities into groups. Using 33 variables, we construct a data set from over 90 publically available sources for 41 different large urban areas and group them into five categories using statistical clustering techniques. Due to missing values, we establish five base clusters using the original data. Then, we use k-nearest neighbor (K-NN) (Kowarik & Templ, 2016) quantile sampling to randomly impute missing values yielding an ensemble of 5,000 clusterings. Using methods from Hornik (2005), we use this ensemble to identify which city cluster memberships are sensitive to changes in values not available in the original data. Our results produce an average hard grouping of cities, or consensus clustering, that is robust to missing data as well as soft clustering identifying the uncertainties associated with a city's cluster membership. This work helps provide JWAC insights into how the 41 large urban areas are similar or different in addition to a framework for future such studies.

We show the resulting average clustering from our analysis in Table ES-1. The average groupings are generally robust to missing data, as we show by identifying the differences between the original data and the imputed data. The cities that are not highlighted represent cities that remain in the same cluster as their baseline grouping,

whereas the cities highlighted in yellow change when the missing data is imputed. The number in brackets next to the highlighted cities indicates the cluster from which the observation moved. We see that all of the movement comes from cluster 4 because Nairobi, Baghdad, and Mexico City each shift to cluster 5 with more destabilized regions.

Table ES-1. Average Clustering Using K-NN Quantile Sampling.

Cluster 1 (10)	Cluster 2 (6)	Cluster 3 (6)	Cluster 4 (9)	Cluster 5 (10)
New York	Beijing	Delhi	Lagos	Cairo
Los Angeles	Tianjin	Mumbai	Jakarta	Kinshasa
Chicago	Shanghai	Kolkata	Bangkok	Karachi
Washington, DC	Chongqing	Bangalore	Manila	Al-Riyadh
Dallas-FW	São Paulo	Hyderabad	Moscow	Tehran
Philadelphia	Rio de Janeiro	Johannesburg	Istanbul	Kabul
San Francisco			Buenos Aires	Dhaka
Boston			Lima	Nairobi[4]
Toronto			Ho Chi Minh City	Baghdad[4]
London				Mexico City[4]

Table ES-1 outlines our consensus (average) clustering after we generate 5,000 samples of data with imputed missing values. The cities in yellow indicate cities that shifted clusters from our original result with missing values. The brackets indicate the original cluster to which they belong.

Cities in Latin America, Africa, the Middle East, and Asia continue to represent the arc of instability (AOI), and clearly cluster together based on their characteristics. Therefore, our results inform the challenges of the AOI. We find that many of the cities within the AOI cluster with others in the AOI, and cities outside the AOI tend to cluster with other cities outside the AOI. We analytically show regions and particular cities that face greater risks due to their shortfalls in the infrastructure, government, economic, and demographic pillars by understanding the cities with which they cluster. As a result, JWAC and other agencies within Department of Defense (DOD) can request more targeted intelligence collection or conduct more effective information operations. Simultaneously, our work shows the need for combatant commands to work in coordination. Each cluster of similar cities contains large urban areas from different combatant commands. And, while we cannot predict with certainty the actual areas where instability or conflict will arise given current data limitations, we can move the discussion forward and identify the areas where effort can increase.

References

Hornik, K. (2005). A clue for cluster ensembles. *Journal of Statistical Software*, *14(12)*. Retrieved from https://www.jstatsoft.org/article/view/v014i12

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software, 74*(7). Retrieved from https://www.jstatsoft.org/article/view/v074i07

United States Army. (2014). *Megacities and the United States Army*. Strategic Studies Group. Retrieved from https://www.army.mil/e2/c/downloads/351235.pdf.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

First, I would like to thank my advisor team, Dr. Whitaker, Dr. Appleget, and Dr. Burks, for their support, patience, and guidance in the development of this work. It has proved to be a challenging and yet rewarding experience that I will **n**ever forget. Your passion and dedication helped to make this possible.

To my parents and sister, your unwavering support for me throughout my graduate education has continued to inspire me. Thank you from the bottom of my heart for all you have done for me.

Finally, to my spouse, Marie, you are my muse, my editor-in-chief, my sounding board, and my motivation. Even from across the country, you have been with me every step of the way. And for that, I thank you most of all. I love you.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

As the U.S. military prepares for the future, senior leaders and analysts alike expect that urban environments will play an increasing role in the operations we conduct. People are migrating to urban areas and the littorals at an increasing rate, which makes understanding the structure and unique challenges of working in these densely populated regions increasingly important. We use a combination of clustering techniques and sensitivity analysis to methodically categorize megacities and large urban areas using a data set constructed using over 90 sources that includes 41 cities throughout the world. This analysis is intended to inform the Joint Warfare Analysis Center (JWAC) on the relationships between megacities and provide insights into the ways in which they are similar or different. It also provides JWAC with a blue print for this type of analysis. Specifically, we give the details and reasoning for how we construct the data set and show how to handle and understand the effects of the inevitable missing values.

A. BACKGROUND

JWAC is responsible for providing senior leaders including the joint staff, combatant commanders, and other key stakeholders with effects-based analysis and targeting options for various critical infrastructure networks. These analyses provide key decision makers with immediate options to support their operational objectives, as outlined in the JWAC mission statement. JWAC uses data from the intelligence community to analyze these networks at all levels and develop the targeting picture. However, the intelligence community does not possess the time or resources to provide JWAC with all of the network detail essential to determining exactly how local networks are structured and which nodes are connected. This limits JWAC’s ability to identify critical nodes and understand precisely the follow-on effects of node interdiction. To reduce the burden on collection and achieve intended results, JWAC focuses its efforts on key components of critical infrastructure, including telecommunications networks and data centers, oil and gas production and distribution, power generation, and the electro-fiber backbone. Even when focusing on these few key networks, however, identifying and synthesizing data proves to be a virtually insurmountable task for regions as complex as megacities.

Currently, over half of the global population lives in large urban areas (Kilcullen, 2013). As population growth continues, analysts predict that most future growth will take place in metropolitan areas (Kilcullen, 2013; United States Army, 2014). Furthermore, Kilcullen (2013) asserts that low-income regions of Asia, Africa, and Latin America will be the primary recipients of the growth as well as the instability that often results from struggles for resources in densely populated, poorly governed areas. These realities increase the likelihood that U.S. forces will be required to operate in densely populated urban terrain in future operations.

Often referred to as megacities, these densely populated urban areas are characterized by a total population of at least 10 million. They may vary widely, however, in terms of infrastructure, demographics, and economic development. With limited resources and intelligence collection capacity, it is virtually impossible for U.S. forces to create a repository of the data necessary to conduct interdependent network analysis on each city we may be required to enter. Moreover, even if perfect information is available, it often requires too much time to collect and process in order to be actionable. Our potential adversaries are dynamic, and as the security environment and risks continue to evolve, it will be beneficial for the military to have general insights and adaptive solutions for managing operations in destabilized regions of the world.

Much of the current literature discusses cities located within the arc of instability, an area of strategic importance to U.S. forces and the focus of this paper. The arc of instability was originally defined in Barnett (2004) and included large portions of Asia and Africa as well as parts of Central and South America. It represents parts of the developing world where competition for resources, terrorism, political instability, and economic inequality make conditions ripe for conflict leading it to become a staple of American foreign policy decisions since the early 2000s (United States Marine Corps, 2015; Barnett, 2004). Barnett (2004) elected to not include India and western China; however, when defined in United States Marine Corps (2015), the arc of instability includes a portion of western China as well as all of India. For the purposes of our study, we will use the arc of instability depicted in United States Marine Corps (2014) and United States Marine Corps (2015), both of which show India being within the arc of instability as shown in Figure 1. In addition to the basic diagram with the arc of instability, Figure 1 also shows all 36 cities that meet the megacity definition in 2015. Those within the arc appear red; those outside the arc appear green.

Figure 1. Arc of Instability. Adapted from United States Marine Corps (2014).

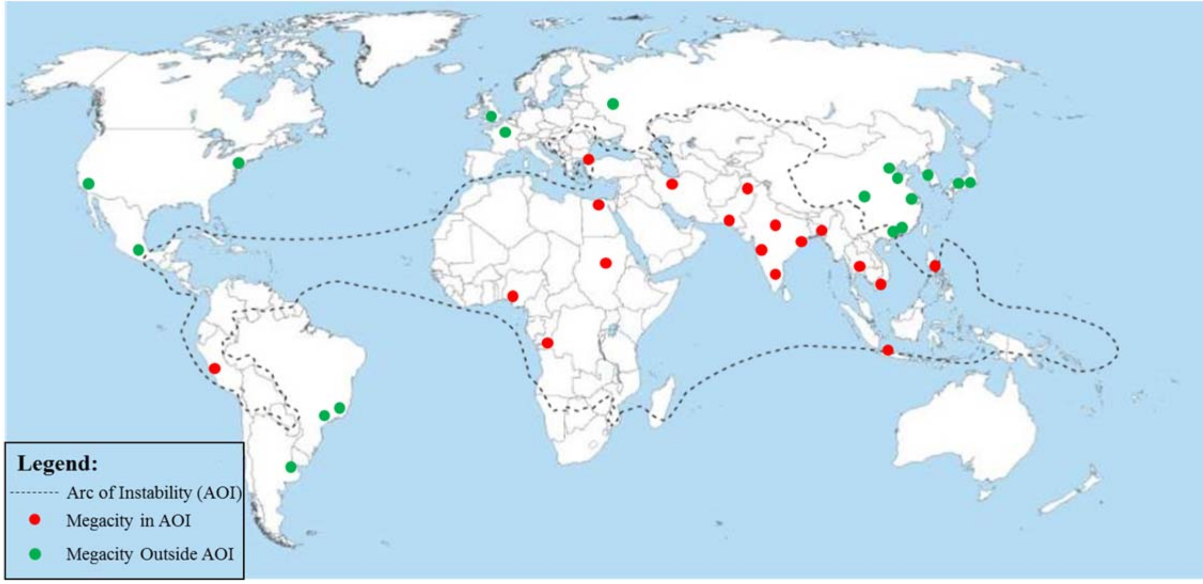


Figure 1 shows a visual depiction of the arc of instability (AOI) taken from a Headquarters Marine Corps Current Operations Brief which was updated in October of 2014. This depiction is a follow-on to the version displayed in Barnett (2004). We overlay the locations of current megacities on top of the map depicted.

Half of all global megacities reside within the arc of instability—and that number is slated to grow. As depicted in Figure 1, the arc of instability encompasses most of Africa and Asia and a small portion of Latin America. According to the United Nations' World Population Prospects (2015), the top nine countries expected to contribute the most to world population growth between 2015 and 2050 include India, Nigeria, Pakistan, Democratic Republic of the Congo, Ethiopia, United Republic of Tanzania, United States, Indonesia and Uganda. And, of these, only the United States does not currently reside within the arc. It is easy to see that the urban environments in the arc of instability will continue to increase in importance in the eyes of defense senior leaders over the next decade, which means that organizations like JWAC must understand their complex dynamics now.

To begin developing its understanding of the structure of various megacities and how their interdependent human and infrastructure networks interact, JWAC has taken an interest in identifying which nodes of these interdependent networks are the most critical and how that knowledge can be exploited to provide sound targeting options for senior leaders and decision makers. Today and in the future, the connectedness of cities and the instant transmission of information through the Internet magnify every action taken by military forces. As a result, commanders are increasingly sensitive to collateral damage and civilian casualties, particularly in densely populated megacities where there are thousands of people per square mile. Simultaneously, the interdependent infrastructure also increases the risk of unintended consequences.

Prime examples of these follow-on consequences occurred in the aftermath of the terrorist attacks on September 11, 2001. In addition to the deaths of 2,996 civilians and the physical destruction of the World Trade Center and a section of the Pentagon, one of the follow-on consequences of the attack was significant network disruption. As identified in Moss and Townsend (2005), the major cellular network carriers reported to the Federal Communications Commission (FCC) a ten-fold increase in call volumes in the moments following the attacks. This level of congestion in the network led to a 92% block rate on New York City’s wireless telephone networks. Simultaneously, the attacks caused extreme disruption to the local transportation network, as well as incoming U.S. air traffic.

While the events of September 11 represent a large-scale complex attack, the concepts also apply to the small-scale targeted strikes that our forces will execute. Hence, in addition to developing an understanding of the interdependencies in these networks, JWAC requires support in determining what risks we face regarding these unintended consequences. While there has been some work studying the effects of disrupting interconnected networks (e.g., Dickenson, 2014), it is virtually impossible to study them without a full data set of network information including network interactions. However, it is possible to gain insight about a particular city based on the cities which have similar networks. Thus, in this thesis, we construct a data set from publically available sources that capture key features of megacity networks and then show how to use this data to find groups of similar megacities.

B. CLUSTERING

In order to produce megacity groups for JWAC, we use statistical clustering methods that are easily implemented and understood. More importantly, they are also well suited to the types of data used to describe megacities network features. Our study focuses on clustering via a partitioning method, which involves using the data to identify which observations are most similar and placing them in the same group or partition. The key to partitioning methods, and indeed most clustering methods, is the choice of how to measure the distance or dissimilarity between observations (in this case cities). This choice is particularly difficult in the data contain mixed numeric and categorical variables with missing values, as is the case with our megacity dataset. We use Gower's distance (Gower, 1971) to measure the inter-point distances between pairs of cities and then cluster those that are closest together using the partitioning around medoids (PAM) algorithm (Kaufman & Rousseuw, 1990). Another key benefit to this type of analysis is that it does not require a specific hypothesis test or design of experiments in order to be effective and provide insights into the data. It is designed to be a non-parametric form of exploratory analysis, which provides us with more flexibility without comprising the integrity of the data. We recommend readers see Kaufman and Rousseeuw (1990) for a review of clustering analysis.

C. OBJECTIVES AND LIMITATIONS

Because we do not have access to the amount and type of data necessary to achieve all of the goals expressed by JWAC, we will use proxies for the detailed network information that will allow us to establish a framework and building blocks for future work. We accomplish this by using meta-data from a selection of large cities around the globe to classify megacities into groups based on their infrastructure, demographic, government, and economic characteristics, which we refer to as "the four pillars." Further, because data uncertainty and missing information are unavoidable, we randomly impute missing values using a k-nearest neighbor (K-NN) method that we then use to construct and ensemble of clusterings. This ensemble is used to determine how sensitive the megacity groupings, or clusters, are to missing information. Through clustering and imputation, we can then provide JWAC with a tool that allows them to gain insight about megacities less understood by identifying key characteristics and comparing them to megacities for which we have more robust data. JWAC can also then use the cluster memberships to conduct case studies using those cities with each cluster that have a repository of detailed network information.

As with any study, the availability and quality of the data limits what we are able to achieve with our analysis. Our data set contains information for 41 cities across the globe, and some cities possess more complete and reliable information than others. An important part of our results are how we decide which data to use and how we mitigate the effects of missing data by using a combination of clustering techniques that are robust to missing values. A sensitivity analysis then enables us to understand the impacts of uncertainty due to those values.

D. THESIS ORGANIZATION

In Chapter II, we address the current literature discussing the megacities and their networks as well as some of the modeling techniques that have been applied to understanding them. We also describe how our work adds to the discussion and deepens our collective understanding of the unique challenges associated with studying megacities. In Chapter III, we describe how we decide which variables to use, how we collect the data, and particular challenges we face in collecting data. We address our methodology in Chapter IV including how we calculate the distances between observations, the algorithm we select, and our data imputation technique. The results and analysis for our study appear in Chapter V where we show our baseline clusters, how the clusters shift when we impute missing values, the variability in the clustering ensembles, and the consensus (average) clustering resulting from our sensitivity analysis. Finally, we use Chapter VI for conclusions and to address where future researchers can further contribute to the study of megacities.

II. LITERATURE REVIEW

While the notion of megacities remains a relatively new concept and literature on tackling its myriad challenges is limited, researchers have conducted substantial research on interdicting and protecting various individual infrastructure networks. Because our focus is not network analysis or optimization, we will not address the matter in significant detail. In this chapter, we will, however, examine the analysis conducted by Dickerson (2014) in order to illustrate the type of research our work seeks to augment and support. We will also explore recent work conducted by the U.S. Marine Corps and U.S. Army as both services begin to look closely at the role megacities will play in their future operations. Finally, we review recent work in grouping like megacities.

A. NETWORK INTERDICTION

Dickenson (2014) uses optimization to examine the interdependence of a power network and its fuel network, where some network nodes are interdependent and some are not. To do this, he develops separate mixed-integer programs for each system, combines them, and then minimizes the total operating cost for the two systems. He uses scenarios to determine the impacts of losing a nuclear power plant. His results indicate that power losses may be spread across any subset of the nodes unless the interdependent nodes are penalized for losing power (Dickenson, 2014). This demonstrates the importance of defending nodes that are known to directly impact other critical infrastructure nodes. However, a key challenge with megacities is that the full network is virtually never known and incorporates many other factors such as water, transportation, telecommunications networks, as well as other elements that cannot be modeled easily using linear or nonlinear programs, such as the impact of the economy or government services on the same system. This provides some of the underlying justification for our approach. By using meta-data for megacities that can capture key features of all of these different attributes, we can better understand how megacities are similar or different.

B. THE FUTURE OPERATING ENVIRONMENT (FOE)

The U.S. Marine Corps' Futures Directorate conducted the 2015 Marine Corps Security Environment Forecast (MCSEF) in order to evaluate the potential operating areas and threats in the years 2030–2045 (United States Marine Corps, 2015). They sought to survey the critical trends and patterns that will shape the FOE and to provide institutional insights into the concepts and capabilities the Marine Corps must develop in order to prepare itself to face those future challenges. To analyze trends, MCSEF examines the changing global demographics, technology developments, competition for resources, stressors on the environment, globalization, governance, and the increasing influence of the urban littorals (United States Marine Corps, 2015). In broad terms, MCSEF discusses increasing movement toward urbanization and littoralization and their respective impacts on the key trends. For our purposes, we will utilize the definition of littoralization outlined in U.S. Marine Corps (2015): a geographic process by which populations and economic activities come together in the littoral environment. The MCSEF goes on to support the claims of Kilcullen (2013) and United States Army (2014) that Africa and Asia will continue to account for a large proportion of the additional urban growth. With the risks of extreme weather and natural disasters in these highly connected urban areas, the complexity of military operations increases, including humanitarian aid and disaster relief missions. Moreover, as a result of U.S. strategic initiatives to rebalance to the Pacific area of operations (AOR) where much of the growth is expected to occur, United States Marine Corps (2015) makes the case that we must be prepared for the full spectrum of military operations in these densely populated cities, including peacekeeping and major combat operations.

The most significant contribution of this study to the problem of military operations in megacities is the Marine Corps' examination of the capabilities and requirements that will continue to be developed in order to achieve our objectives in the FOE. The United States will continue to place greater emphasis on using unmanned and autonomous system technologies as well as precision guided munitions because of the increased risk associated with operating in an urban environment. Civilian casualties and collateral damage are far more likely in densely populated urban centers, and with the widespread use of social media, any tactical mishaps that cause unintended consequences can have strategic level impact. This underscores the importance of deepening our understanding of the structure and interdependence of megacities, as well as why JWAC has a keen interest in the matter.

Unfortunately, like much of the current work concerning megacities, United States Marine Corps (2015) does not provide quantitative data or analysis beyond broad global trends and patterns. While identifying these trends provides senior leaders with context, commanders need information specific to their regional areas of operations that will support decision making. Quantitative analysis helps to shape the operating picture for commanders by providing key insights and recommendations backed by data. We show that our quantitative analysis augments this work by providing insights into the structure and nature of various large urban areas. By creating megacity clusters and identifying the variables that draw cities close to one another, we can gain insight about those cities with less information.

C. GROUPING SIMILAR MEGACITIES

In one of the first attempts to apply academic rigor to studying megacities, the U.S. Army's Strategic Studies Group (United States Army, 2014) conducted an analysis of megacity dynamics and challenges by applying qualitative methods to six different case studies including: Dhaka, Bangladesh; Lagos, Nigeria; New York City, New York; Bangkok, Thailand; Rio de Janeiro, Brazil; and São Paulo, Brazil. In the first four case studies, the researchers were able to conduct field work, whereas they only used a virtual case study method in Rio de Janeiro and S*ã*o Paulo. Specifically, they utilize a systems-theory approach to analyze the megacities holistically because of their core belief that the whole of the system must be appreciated in order to understand its various subcomponents.

They follow this technique to establish two key systems and numerous characteristics associated with each system. The first system, "City Characteristics," addresses the physical, economic, and social infrastructure of the cities to establish baseline context (United States Army, 2014). Then, the United States Army (2014) adds second system, “Dynamics of Instability and Capacity,” which identifies the impact of evolving demographic features as well as internal and external threats on the resilience of the cities. For example, the combination of substantial urban migration and ethnic separation along with high levels of inequality may reduce the ability of a megacity to fight the destabilizing effects of hostile actors.

As a result of their analysis, they produce three qualitative cluster groupings for the six megacities in their case study. The first cluster includes highly integrated cities with centralized and formal systems, high quality infrastructure, and regulated flow capacity (United States Army, 2014). The second cluster, moderately integrated cities, contain a mix of formal and informal systems, mixed quality infrastructure, and self-regulated flow capacity (United States Army, 2014). Finally, loosely integrated cities are those which have decentralized and informal systems, poor quality infrastructure, and unregulated flow capacity (United States Army, 2014). As part of their findings, they conclude the cities from their case study can be grouped as shown in Table 1.

Table 1. Megacity Case Study Groupings. Adapted from United States Army (2014).

Highly Integrated	Moderately Integrated	Loosely Integrated
New York City, New York	Bangkok, Thailand Rio de Janeiro, Brazil São Paulo, Brazil	Lagos, Nigeria Dhaka, Bangladesh

This table illustrates how the six cities studied in United States Army (2014) break into clusters based on the qualitative characteristics identified by the Strategic Studies Group.

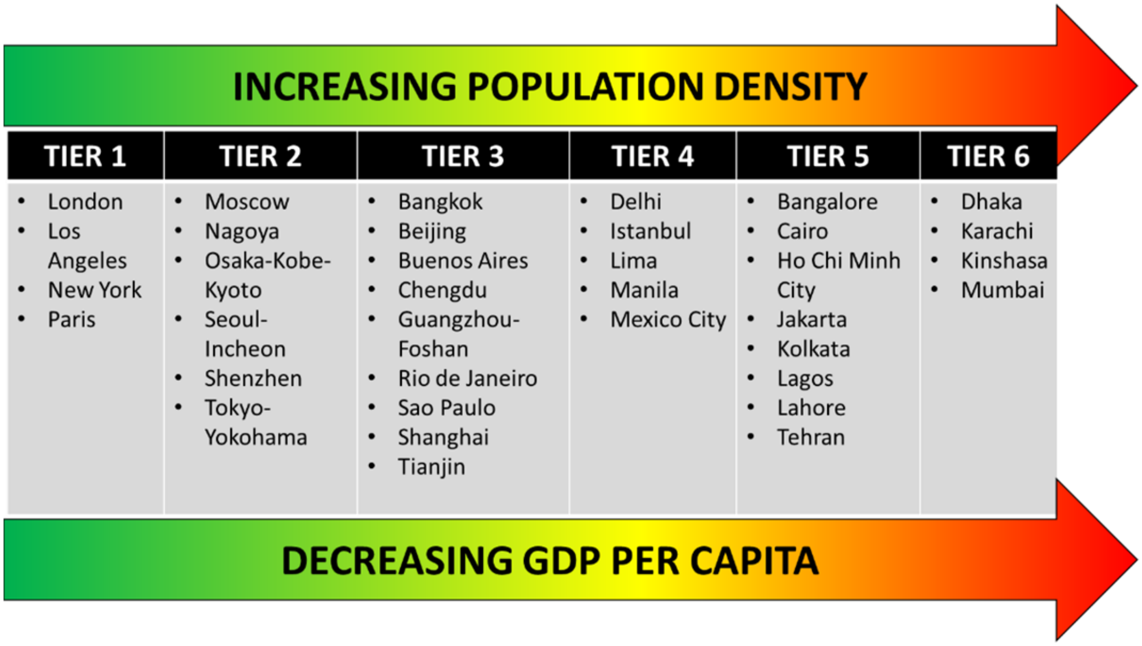
As might be expected, New York is a highly integrated, complex, and densely connected city. However, without a deep understanding of the other five cities studied, one might not expect the other cities to be grouped as they are in Table 1. Notably, one might expect Lagos to be a moderately integrated city and Bangkok to be a loosely integrated city. Additionally, the evidence United States Army (2014) provides for their decision in each case study demonstrates some academic rigor. They address key data points for some of the cities pertaining to population, gross domestic product (GDP) growth, and infrastructure such as the estimated number of buildings. They also incorporate military assessments regarding the U.S. Army's ability to adapt to such a city. Furthermore, their case study method provides greater depth than the broad discussion of megacities utilized in United States Marine Corps (2015).

However, United States Army (2014) is limited in terms of quantitative depth. They do not clearly identify the actual infrastructure characteristics that led them to place the cities in specific categories such as the number of and flow rate through international airports, city access to mass transportation, amount of electricity production or number of power plants, among other aspects the would more clearly articulate what it means to have a highly integrated city. Furthermore, field work requires significant time and resources. In fact, the team was not expected to visit Mexico City, their seventh case study, until July of 2014, which was after the study publication date of June 2014. Clearly, efforts to expand the case study to incorporate each of the 36 current megacities in the world would prove burdensome. This is where our study adds value to the discussion regarding U.S. military's understanding of megacities. We collect city level data pertaining to the four pillars (infrastructure, government, demography, and economy), and we apply quantitative analysis to support the work completed by the strategic studies group.

Sapol (2016) provides a basis for analyzing megacities with quantitative techniques in order to influence how the U.S. military aligns forces. The current paradigm is tied to the six major regions aligned with geographic combatant commands (COCOM), but his study indicates the potential for shifting to a megacity focus in order to potentially reduce the number of political, military, economic, social, information, infrastructure, physical environment, and time (PMESII-PT) variables that must be evaluated. We recommend readers review United States Army (2013) for additional details regarding how the military develops PMSEII-PT information. One advantage of this approach is that combatant commanders can locate a megacity of interest in a cluster of similar, but potentially geographically disparate, cities, and seek recommendations from fellow commanders operating within those cities. Clustering thus induces combatant commanders to look outside their own AOR and instead to leaders operating in more relevant environments, resulting in efficiencies in information sharing and strategic advantages in combat.

In order to develop the clusters, Sapol (2016) collects the population density estimates and GDP per capita for each of the 36 megacities from the time of the study to serve as proxies for key features within a city. Sapol (2016) argues that GDP per capita is indicative of military spending, economic status, information systems, and infrastructure, whereas population density is indicative of infrastructure and the physical environment. By simplifying his analysis to these two high-level variables, he is able to easily collect, display, and explain the data to senior leaders. As a result, he produces six tiers of megacities such that cities in tier 1 possessed a high GDP per capita and low population density, and cities in tier 6 were characterized by low GDP per capita and high population density. Cities in the middle tiers are placed on a sliding scale between the tier 1 and tier 6 result. The groupings for the six tiers can be seen in Figure 2.

Figure 2. Megacities from Sapol (2016) Clustered by GDP Per Capita and Population Density.



TIER 1	TIER 2	TIER 3	TIER 4	TIER 5	TIER 6
• London • Los Angeles • New York • Paris	• Moscow • Nagoya • Osaka-Kobe-Kyoto • Seoul-Incheon • Shenzhen • Tokyo-Yokohama	• Bangkok • Beijing • Buenos Aires • Chengdu • Guangzhou-Foshan • Rio de Janeiro • Sao Paulo • Shanghai • Tianjin	• Delhi • Istanbul • Lima • Manila • Mexico City	• Bangalore • Cairo • Ho Chi Minh City • Jakarta • Kolkata • Lagos • Lahore • Tehran	• Dhaka • Karachi • Kinshasa • Mumbai

As pictured, the cities with the highest per capita GDP and lowest population density make up the tier 1 cities. As the population density increases and per capital GDP decreases, the tiers transition until reaching the highest density and lowest per capita GDP cities in tier 6.

When we compare the results of this study to the results of the study conducted by the Army Strategic Studies Group (United States Army, 2014), we can see that some of their claims support each other. Colloquially termed “western” cities such as New York City fall into the highly integrated or tier 1 category, cities such as Rio de Janeiro or Bangkok fall into the middle tier in each study, and developing world cities such as Dhaka and Lagos fall into the or tier 5 and 6 categories. These results demonstrate the value of using quantitative data to conduct analysis in Sapol (2016). For less time and cost, he is able to produce results similar to those in United States Army (2014) while simultaneously increasing the scope of the work by incorporating all other megacities across the globe.

While Sapol (2016) provides an initial analytical examination of the problem of grouping megacities, it does not provide depth and makes a critical assumption regarding its two variables. The Army Strategic Studies Group (United States Army, 2014) indicates that each megacity possesses its own characteristics and attributes. Hence, no two cities will exhibit exactly the same make-up across the four pillar. This makes reducing the many factors that make up those four major areas to GDP per capita and population density as outlined in Sapol (2016) unrealistic. Moreover, the six tiers based on these two variables do not provide military commanders with actionable information. For example, Chicago and Los Angeles would both be considered tier 1 cities using methods from Sapol (2016). However, they are quite different in terms of critical infrastructure and distribution of emergency services. Defined by their respective metropolitan statistical areas (MSA), Los Angeles has access to a major seaport for transporting goods into the country, but Chicago has a larger number of international and regional airports. Moreover, Chicago has a smaller population but consumes a larger amount of water per day and has a more robust public transportation system. Hence, a more robust solution might dictate that Los Angeles and Chicago are both tier 1 cities as outlined in Sapol (2016), but they split into separate sub-groups due to their differences in actual infrastructure, government, economic, and demographic attributes.

Our work furthers the discussion on addressing the nature of various large urban areas by constructing a more comprehensive data set and applying sensitivity analysis to determine where shifts in cluster membership may occur given data uncertainty. This same sort of sensitivity analysis can also be a tool used to make observations regarding how the cluster memberships change when new data is added. For example, a city like Lagos may group more closely with a different set of cities than in Sapol (2016) if they add more robust infrastructure, demographic features change with time, or more reliable data becomes available.

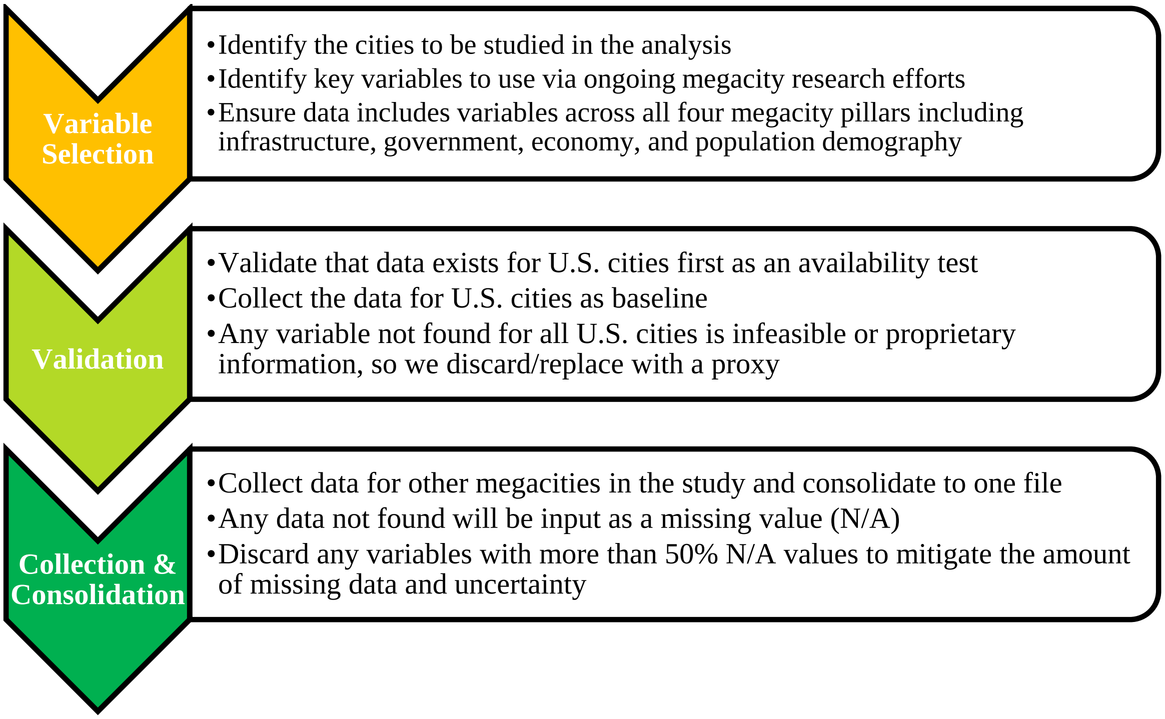
III. VARIABLE SELECTION AND DATA COLLECTION

We construct the data set for this analysis using a combination of infrastructure, economic, government, and demographic information for 41 different large cities across the globe. Because there is no single database that addresses all of these categories for large urban areas, we need to accomplish two things before we built the data set. First, we need to identify which variables would provide enough information about a city to make an informed and reasonable assessment about its character. Second, we need to assess the feasibility of obtaining that information for cities in a list of candidate cities before beginning the search. In order to identify exactly which variables to incorporate, we begin with concepts identified in United States Army (2014), United States Marine Corps (2015), Sapol (2016), and Kilcullen (2012). Each of these documents include some insights into the type of information required for this type of study and allow us to capture this information into a set of 33 variables extracted from over 90 public sources. Specifically, we group information into our four pillars including infrastructure, government, economy, and demography.

A. DATA SELECTION

For infrastructure, we examine electricity generation/capacity, quantities of water sources by type (ground/rivers and reservoirs/desalination) and daily water distribution, the number of miles of road and rail within the cities, the number of regional and international airports, whether the city has a subway system as the rail system, whether the city contains a seaport as well as the flow through the seaport, the percent of the population with telephone access, the percent of the population with Internet access, and percent of the population with access to sanitation. For government services, we construct variables for whether the city is a state/provincial or national capital, hospitals, fire stations, police stations, and military presence. As indicators of economic health, we add the gross metropolitan product (GMP) or gross regional domestic product (GRDP), household income, the poverty rate, and education. Then, to evaluate the structure and composition of the population, we include the literacy rate, the percentage of young people, the percentage of elderly people, the primary religion practiced by the population, number of people per household, and the population density. In the following sections, we detail the underlying justification for our variable selections as well as key variables we were unable to include but might improve future studies on the subject. A summary of our data collection process can be seen in Figure 3.

Figure 3. Data Collection and Validation Process.



Variable Selection

- Identify the cities to be studied in the analysis
- Identify key variables to use via ongoing megacity research efforts
- Ensure data includes variables across all four megacity pillars including infrastructure, government, economy, and population demography

Validation

- Validate that data exists for U.S. cities first as an availability test
- Collect the data for U.S. cities as baseline
- Any variable not found for all U.S. cities is infeasible or proprietary information, so we discard/replace with a proxy

Collection & Consolidation

- Collect data for other megacities in the study and consolidate to one file
- Any data not found will be input as a missing value (N/A)
- Discard any variables with more than 50% N/A values to mitigate the amount of missing data and uncertainty

1. Infrastructure

In order to sustain businesses, government activities, and residential environments electricity is essential to the industrialized world. As such, power grids are also at risk of attack during conflicts or disruption during natural disasters. Furthermore, in the current and future operating environment, these attacks and disruptions will often occur without kinetic military strike packages such as guided missiles and bombs but rather through non-kinetic attacks that shut all or portions of the network down. Specifically, Smith (2016) discusses the growing fears among government officials regarding cyber intrusions. As part of the report on Russian intrusions into the U.S. and Ukrainian power grids, former homeland security advisor, Frank Cilluffo, raises concerns about the implications of these actions on national security (Smith, 2016). In short, the ability of adversaries to conduct attacks on critical infrastructure without firing a shot or risking lives changes the operational picture dramatically. Airports, seaports, hospitals, manufacturing plants, and public transportation all rely upon access to electrical power which suggests that major disruptions can have far-reaching impacts. The Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) from the United States Department of Homeland Security (DHS) also addresses some of this impact in a report regarding other suspected Russian power grid attacks against Ukrainian infrastructure in December of 2015. Specifically, ICS-CERT (2016) reports that the attack resulted in loss of power to 225,000 customers, while striking multiple facilities in the area. And, a British Broadcasting Corporation (BBC) News report identifies that another power outage in 2016 resulted in the loss of power to approximately one fifth of the total power consumption in Kiev, likely due to a cyber-attack (BBC, 2017). These outages indicate a growing capability and significant risk for densely populated urban areas like megacities. As a result, we initially sought to include the number of power plants in a city by type of power source as well as the total generating/capacity in megawatts (MW). However, detailed data pertaining to the number of power plants by power source in the international cities proved to be unavailable, so we only include the total installed capacity/generation in this study.

In addition to power, water represents another critical utility that influences life in a megacity. Rivers and dams provide hydroelectric power, as well as sources for drinking water and sanitation. Tortajada et al. (2006) present some of the key challenges that large urban areas face with regard to water sources, supply, demand, and distribution. Many developing megacities such as São Paulo, Mexico City, Dhaka, and Al-Riyadh face challenges producing enough water to distribute through their water network due to inefficiencies and poor infrastructure. They further assess that this often results in the supply not being able to meet water demand. In addition to the risk of kinetic military strikes and non-kinetic attacks like cyber intrusions, many megacities risk disruption and contamination to their water supply during natural disasters. As outlined by Moroney et al. (2013), the Department of Defense (DOD) plays a significant role in the humanitarian assistance and disaster relief (HA/DR) mission due to our budgetary resources, robust logistical capabilities, and consistently forward deployed status. Moroney et al. (2013) go on to outline several HA/DR missions where the U.S. military provides various types of support, including water and sanitation provisions. Hence, it is important for us to capture indicators of cities' current water infrastructure because we expect that cities with less robust infrastructure and poor ability to meet water demand face greater risks in the event of disasters such as Hurricane Sandy in Haiti or typhoons in the Philippines, in addition to military strike outages. We had difficulty accumulating detailed information regarding the number of water treatment facilities and output efficiency. This led us to use the number of water sources by type including rivers/lakes/reservoirs, ground water, and water desalination, as well as the daily water supply to the city in millions of gallons as proxy variables. We expect that the water source information is indicative of the infrastructure and susceptibility to contamination, and the daily supply indicates a capacity and flow to the population.

Kilcullen (2013), provides a clear example of the ways terrorists can exploit access to seaports and the broader water network while improving their ability to hide amongst the populace in a dense city like Mumbai, one of the 41 megacities identified in our study. Specifically, he discusses a ten-man of Pakistanis from the terrorist organization, Lashkar-e-Taiba, using a ship sailing from the Karachi's harbor to the Port of Mumbai in order to land and hide amongst the populace before conducting an attack (Kilcullen, 2013). This underscores one important aspect of incorporating whether a city has a seaport, but there are also others. In addition, we know that navies require access to seaports to conduct re-supply. And, a seaport is also an indicator of the city's economy. As reported by the United Nations' (UN) International Maritime Organization (IMO), 90% of global trade volume moves by sea which suggests that any impacts to the flow of goods or services in or out of a seaport will have rippling effects on the economy (United Nations, IMO, 2017). From a doctrinal perspective, this helps to explain the importance of a naval blockade as a tactic for influencing adversaries. As a result, in addition to capturing whether a city has a seaport, we also found it beneficial to capture economic flow through the seaport as an indicator of economic value. We accomplished this by adding the annual number of twenty-foot equivalent units (TEUs) and total throughput in millions of metric tons. This allows us to identify availability as well as flow in order to make considerations for the impact of disruption through disasters or strikes.

In addition to transportation by sea, airports represent a growing part of the transportation of goods and services as well as people. In the event of conflicts, airfields can be utilized to transport military supplies or struck as targets. At the turn of the century (2000), the World Bank reports that approximately 118,257 million ton-km of freight moved across the globe via air. And, by 2015, the World Bank reports this number reached 188,000 million ton-km which represents a 59% increase over that period (World Bank, 2017a). More impressively, the number of passengers transported more than doubled, increasing from 1.67 billion to 3.44 billion in the same period (World Bank, 2017b). This illustrates the increasing importance of airports and aircraft to the overall infrastructure of cities. As a result, we sought to capture an indicator for the amount of aircraft-capable infrastructure present in each city. To accomplish this differentiation, we separate airports by whether they are local or international hubs. This led us to include the number of regional airports and the number of international airports in each city. While we were not able to include an indicator for flow at the individual city level, total passengers and freight transported via air may provide interesting insights into capacity for each of these large urban areas.

Obviously, ground transportation networks also play a pivotal role in characterizing the infrastructure of a megacity. The military implications of ground transportation networks are significant and highlight many ongoing challenges in large urban areas. In terms of road networks, the local populace is a concern due to the significant increase in the number of cars on roads that has occurred in the 20th and 21st centuries. Additionally, highways provide another means of moving military supplies and equipment as well as freight away from ports. For example, the Port of New York-New Jersey's website provides estimated transit times to get good transport inland from international locations such as Mumbai, Singapore, and Rotterdam through the available road networks. Kilcullen (2013) also addresses the issue of road networks in an example about the Mungiki gang in Nairobi, Kenya, using food transportation routes to extort money from the operators. Even in cities with robust mass transit systems like New York City, traffic congestion remains a significant challenge and would be further disrupted in the event of disasters or military operations.

In one of the earliest research efforts regarding the military implications of ground networks, Harris and Ross (1955) studied the optimal ways of blocking supplies moving along the Soviet Union’s rail network to Europe in a limited number of attacks. Additionally, most militaries possess heavy equipment such as tanks that are difficult to transport via trucks due to a limited number of heavy equipment transports (HET) and are often transported via rail on the ground, which further supports the strategic importance of understanding the rail network associated with a megacity. And, we also recognize that subway systems are increasing in popularity due to their lower impact of surface infrastructure and ability to move people through an urban environment efficiently. As a result, we find it important to provide an indication of that level of infrastructure, particularly given subway systems are less common in developing world cities. The presence of a subway may be indicating something about the development of the city as well as their economy due to the significant investment costs associated with the development of a subway. For example, Singapore has produced one of the most advanced subway systems in the world for a total cost of approximately $5 billion, which provided approximately 22 miles of track and 28 different stations (Lepeska, 2011). As a result, we include the number of miles of road, the number of miles of rail, and a binary indicator for whether the city has a subway system. These parameters serve as proxies for our original intent to identify the number of transportation choke-points for each city, which would have indicated the capacity of the network to flow goods and people in and out of the city during disruptions.

Telecommunications networks and Internet access both contribute measurably to the infrastructure of a city while simultaneously influencing security. United States Marine Corps (2015) argues the ubiquity of telecommunications is decentralizing military, economic, political, and social power to the individual level, allowing small groups to harness mass communications to challenge the authority of the state. We can look to countries like Afghanistan as examples of this trend. In these developing countries, mobile telephone networks enable locals to communicate and conduct business while aiding nefarious organizations like the Taliban and Al Qaeda to coordinate efforts and execute information operations (IO) campaigns against their adversaries (the government and outside actors). United States Army (2014) provides another example of this reality when discussing the ability of gangs using mobile phones in São Paulo to manage illicit drug networks in prisons and informal communities known as favelas. From a combating terrorism perspective, we have also seen the rise of the Islamic State of Iraq and Syria (ISIS) expand the presence of terrorist activities in the online domain. They have utilized the Internet through YouTube, Twitter, and other social media outlets to conduct recruiting, messaging, and publicizing their attacks, which U.S. forces must be prepared to combat in the FOE. Thus, we find it critical to incorporate indicators for the availability of and access to telecommunications and Internet networks. Unfortunately, data pertaining to hard infrastructure was unobtainable, so we include the percent of the population with access to telecommunications (mobile or landline) and the percent of the population with access to the Internet as proxy variables for capacity and network infrastructure.

2. Government Services

Government infrastructure plays a role at all levels of a society, to include the city level. In particular, cities that serve as capitals contain additional civil servants and government buildings to facilitate government operations. For example, in a democracy or republic, we can reasonably expect there to be buildings for the parliament or congress, the courts, the president or prime minister, and all federal agencies at the national level. Moreover, even in an absolute monarchy such as Saudi Arabia, ministries and departments exist that require facilities in addition to a large palace for the king and his family. And, at the state level the same types of government infrastructure are required to support a governor and their staff. Unfortunately, we do not have a sufficient proxy variable to address the details of the infrastructure quality or mass relative to the rest of the city, but we add binary variables for whether the cities are capitals in order to account for the reality that we know some additional infrastructure exists for these cities.

The government sector also provides services to the people in their municipalities such as fire fighters, police, military, and some public hospitals. By accounting for the availability of these resources, we improve our ability to understand potential shortfalls during military operations in megacities, HA/DR missions in particular. Felix and Wong (2015) show that availability and response time of emergency services can have far reaching impacts on megacities. For example, the emergency services available in New York City during 9/11 were able to respond quickly and efficiently because they had robust infrastructure and personnel to support the effort. Hospitals had the ability to treat injuries, the fire fighter network effectively controlled the spread of fires and provided support to evacuation, and the police aided in cordoning the area, providing security, and facilitating evacuation (Felix and Wong, 2015). Simultaneously, these services also illuminate the importance of alternative support access. In the case of Hurricane Katrina, the U.S. military provided medical care as well as evacuation support for thousands of people according to Wombell (2009), which shows the additional capability provided by having military support readily available. As a result, we elect to include the number of fire stations, police stations, hospitals, and military bases in the local area as proxies for coverage because we expect that cities with more emergency services available will be more resilient to disruptions than those with less. We recognize this does not indicate the ability of emergency services to operate under their own power in the event of outages, but this provides a building block for future research.

While we were not able to include controls for levels of corruption or government systems strength, we recognize that these are important features of the government pillar and recommend them as data additions in future work. Kilcullen (2013) and United States Army (2014) discuss this reality exists for many megacities in the developing world, to include São Paulo. They make the case that corruption and organized crime both have significant influence over cities in the developing world and may undermine access to emergency services despite a large amount being located with the metropolitan area. Researchers can also look to examine relative crimes rates, fire responses, and hospital beds per 1000 people because of their ability to examine effectiveness and capacity of emergency services in addition to availability. This data is difficult to find and consolidate for cities, which impeded our ability to include that information in this study.

3. Economic Health

Sapol (2016) gave economic power for megacities an explicit role in his analysis by incorporating GDP per capita and assuming it indicates other aspects of city information as we addressed in Chapter II. Because we explicitly control for factors such as infrastructure and information systems, we view GDP as a primarily an indicator of the economic health of the city. For example, Dobbs et al. (2011) make the claim that 600 cities accounted for 60% of global GDP in 2011. This suggests that much of the economic power of the globe resides in large urban areas which makes sense, but they do not illuminate what that distribution looks like among those 600 cities. On the extreme end, New York City's GDP was approximately $1.13 trillion in 2014 which made it rank 13th in the world among countries according to United States Army (2014). As a result, we make the assumption that the largest cities tend to contribute the most to overall GDP and include it as a parameter for each city in billions of U.S. dollars. To account for the average economic power of the family level, we consider the average household income. We include the individual wealth because we expect some cities in the developing world to have relatively high city level GDP but still low average household incomes, whereas cities like New York City and London both have large GDPs and high average household incomes. To capture this trend, we include average income as a variable in dollars.

Another indicator of economic health is the number of people living in poverty relative to the entire population, which may be driving down the average household income. From an economic perspective, obviously we expect cities with a lower incidence of poverty to have healthier economies. Kilcullen (2013) and United States Army (2014) often reference the risks that income inequality and high poverty rates pose to security in large urban areas. They argue that highly impoverished slums tend to be breeding grounds for high crime, which has important implications for potential military operations in a megacity. Thus, we incorporate each city's poverty rate as an indicator of this reality. In identifying this as a parameter, we also note that poverty is relative to individual national standards and income. Hence, the poverty rates identified in our study are relative to the country in which the cities reside.

The economic health of a city also depends upon the education of the population. In general, different jobs require different skill sets, and a highly industrialized economy requires more highly educated people in order to conduct daily business. Hanushek and Woessmann (2010) discuss the importance of a highly educated workforce at the country level, arguing that improving education can have far reaching impacts on the overall growth of an economy. While our study focuses on the city level vice the country level, many of the same concepts still apply. Automation is a growing part of the industrialized global economy, which suggests that laborers require additional technical and comprehension skills in order to work with the new technologies. United States Marine Corps (2015) agrees that technological advancements will continue to drive economies into the future, which supports the claims that producing an educated workforce adds to both security and prosperity. Because we are not able to obtain the data associated with education quality at the city level outlined in Hanushek and Woessmann (2010), we control for the education of the populace by including the percentage of the population who completed secondary school or high school as well as the percentage of the population who completed a bachelor's degree or equivalent and higher. Future work can also look to compiling education infrastructure information for each city as a signal for education. This might include the number of high schools, universities, and/or libraries in each city. Simultaneously, this may also provide better insight into the common hard infrastructure in each city.

While some may argue that literacy rates provide a similar signal to that provided by education, we refer back to the Hansushek and Woessmann (2010) argument that quality of education is more important to addressing economic development than simply level of education achieved. Specifically, they show evidence that even countries improving in educational level output are still below average in literacy, which suggests that education level does not always perfectly align with performance. Because of this tension between performance and education level, we include literacy rates for cities as a proxy variable for the quality of education.

4. Population Demography

The most common demographic feature among megacities discussed in the literature is population density. Sapol (2016), United States Army (2014), and Kilcullen (2013) all discuss the challenges that densely populated areas pose. Kilcullen (2013) discusses this in the context of allowing non-state actors to hide among the populace, and United States Army (2014) considers density a factor that contributes to environmental vulnerability and resource competition. Finally, because Sapol (2016) also incorporates population density directly into his clustering model, we believe these ideas guide our understanding of the military implications of operating in a more densely populated city vice a less densely populated city. Strikes risk more civilian casualties, natural disasters impact a larger portion of the population in a smaller area, and non-state actors have improved freedom of movement, particularly when they have influence over locals as discussed in Kilcullen (2013). Therefore, we include population density for each city in the form of people per square mile.

The structure and composition of a population are inherently related to infrastructure, the economy, and government services. As identified in United States Marine Corps (2015), the size of the working age population has ramifications for military, economic, political, and social power because they encompass the military-age population, academics and innovators, and the general workforce. Furthermore, in most regions of the world, this age group makes up the vast majority of the populace. However, there is also an interesting relationship between the number of young people in a population and the number of elderly people in a population depending on whether they come from a highly developed part of the world or developing part of the world. In general, populations in highly developed parts of the world have higher life expectancies and a larger percentage of older people, whereas populations in the developing world tend to have a larger percentage of young people.

Nikolova (2016) and Patna (2013) both address aging populations and changing demographics and the impacts of those changes on the broader society. Moreover, the Central Intelligence Agency (CIA) Factbook shows that countries like Afghanistan, Nigeria, and The Democratic Republic of Congo have a median age of approximately 18 years old, whereas countries like the U.S. and U.K. have median ages closer to around 40 (Central Intelligence Agency, 2017). Considering these dynamics, we sought to include indicators for the age demography by identifying the percentage of people 18–19 years old and younger as well as the percentage of the city population ages 65 and older.

We also expect family size varies by region due to differences in cultural norms and expectations. Developed countries tend to have smaller household sizes, and the developing world tends to have larger household sizes. According to a report published by Nakono, an industry research company, we can see that this is accurate on the country level. Saudi Arabia, Philippines, and Egypt, had estimated household sizes of 5.79, 4.5, and 4.35 in 2012, respectively. However, the China, the U.S., and U.K. had relatively smaller household sizes of 3.03, 2.63, and 2.39 in 2012, respectively (TekCarta, 2017). Countries like Afghanistan serve as a future example of this because of family structure and responsibilities. Often, multiple generations will live within the same household in order to gain wealth and spread the work-load associated with family chores. We expect that a similar dynamic may exist at the city level, so we include average household size in our study.

From a cultural perspective, religious affiliation, or lack thereof, play a role in the way people interact with the world around them. While the United States has a mix of religious affiliations, Christianity and its values permeate through American culture even though it is a society with the separation of church and state. According to a Pew Research Center survey, they estimate that approximately 70.6% U.S. citizens describe themselves as Christians. And moreover, only 5.9% of the remaining population affiliates with a specific non-Christian religion such as Judaism or Islam (Pew Research Center, 2017). In the United Kingdom, Christianity proved to be less dominate with 59% of the population affiliating with Christianity, and approximately 25% being unaffiliated (United Kingdom, 2011). In contrast, many countries in the Southwest Asia and Africa affiliate mostly with Islam and simultaneously combine church and state to make Islam the official religion of the country, while Hinduism dominates in India. These dynamics are important to capture in identifying the nature of a city because of the impacts on society and infrastructure. Additionally, from a military perspective, religious structures are generally protected against military strikes. We observed this frequently during the conduct of counter-insurgency (COIN) operations in Iraq where terrorists willingly placed innocent civilians in harm’s way in order to avoid confronting U.S. forces. As a result, it is important to have an understanding of the types of structures we can expect to be most common in a given area, whether they are churches, mosques, or temples.

B. DATA COLLECTION AND VALIDATION

After identifying all of our variables, we collect the data for the U.S. cities in our data set first in order to establish a baseline for the other cities. We assume that any data we are unable to find for U.S. cities would likely be unavailable for foreign cities, so we drop those variables. We also include eight of the top 10 largest cities in the U.S. by population although they are not all megacities. These are New York City, Los Angeles, Chicago, Dallas-Fort Worth, Washington, D.C., San Francisco, Philadelphia, and Boston. We define city borders based on their respective metropolitan statistical areas (MSA) as defined by the Census Bureau because the megacity construct examines the broader metropolitan area rather than the strict city limits. We do this for two reasons: (1) it increases the number of cities in the collection, and (2) it allows for the possibility that large cities in the developing world may be more similar in their infrastructure and economic power to large (but not mega) cities in the U.S. This might, for example, result in clusters that include a large developing world city like Jakarta or Mexico City and a much smaller city in the U.S. like Boston or San Francisco. Once we obtain the necessary data for the cities in the U.S., we search for data regarding the other 33 cities of our study. We select cities from each combatant command that are reasonably large, come from a diverse set of countries, and have some level of strategic importance. For example, Baghdad has strategic relevance due to the fight against ISIS, and Kabul remains an important part of U.S. foreign policy in 2017. Table 2 lists all the selected cities and their population sizes according the Demographia Largest Urban Areas population estimate for 2015 (Demographia, 2016).

Table 2. List of Selected Cities. Adapted from Demographia (2016).

COMBATANT COMMAND	Rank By Population	Geography	Urban Area	Population Estimate
PACOM	2	Indonesia	Jakarta	31,320,000
PACOM	3	India	Delhi, DL-UP-HR	25,735,000
PACOM	5	Philippines	Manila	22,930,000
PACOM	6	India	Mumbai, MH	22,885,000
CENTCOM	7	Pakistan	Karachi	22,825,000
PACOM	8	China	Shanghai,SHG-JS-ZJ	22,685,000
NORTHCOM	9	United States	New York, NY-NJ-CT	20,685,000
SOUTHCOM	10	Brazil	São Paulo	20,605,000
PACOM	11	China	Beijing,BJ-HEB	20,390,000
NORTHCOM	12	Mexico	Mexico City	20,230,000
EUCOM	15	Russia	Moscow	16,570,000
PACOM	16	Bangladesh	Dhaka	16,235,000
CENTCOM	17	Egypt	Cairo	15,910,000
PACOM	18	Thailand	Bangkok	15,315,000
NORTHCOM	19	United States	Los Angeles, CA	15,135,000
PACOM	20	India	Kolkata, WB	14,810,000
SOUTHCOM	21	Argentina	Buenos Aires	14,280,000
CENTCOM	22	Iran	Tehran	13,670,000
EUCOM	23	Turkey	Istanbul	13,520,000
AFRICOM	24	Nigeria	Lagos	12,830,000
SOUTHCOM	26	Brazil	Rio de Janeiro	11,815,000
AFRICOM	27	Congo (Dem. Rep.)	Kinshasa	11,380,000
PACOM	28	China	Tianjin,TJ	11,260,000
SOUTHCOM	29	Peru	Lima	10,950,000
EUCOM	33	United Kingdom	London	10,350,000
PACOM	34	India	Bangalore, KA	10,165,000
PACOM	35	Viet Nam	Ho Chi Minh City	10,075,000
NORTHCOM	39	United States	Chicago, IL-IN-WI	9,185,000
AFRICOM	40	South Africa	Johannesburg-East Rand	8,655,000
PACOM	43	India	Hyderabad, TL	7,750,000
PACOM	47	China	Chongqing,CQ	7,440,000
CENTCOM	54	Iraq	Baghdad	6,790,000
NORTHCOM	56	Canada	Toronto, ON	6,550,000
NORTHCOM	58	United States	Dallas-Fort Worth, TX	6,280,000
NORTHCOM	65	United States	San Francisco-San Jose, CA	5,955,000
CENTCOM	66	Saudi Arabia	Riyadh	5,845,000
NORTHCOM	73	United States	Philadelphia, PA-NJ-DE-MD	5,595,000
NORTHCOM	80	United States	Washington, DC-VA-MD	4,950,000
AFRICOM	81	Kenya	Nairobi	4,930,000
NORTHCOM	91	United States	Boston, MA-NH-RI	4,490,000
CENTCOM	121	Afghanistan	Kabul	3,650,000

Table 2 contains each city, as defined by their metropolitan area, used to conduct this study as well as the COCOM responsible for operations in that area, their world population rank, and estimated population as of 2015. Recall, cities with a population greater than 10 million are considered “megacities.”

C. DATA SUMMARY

The data set for this study incorporates a total of 33 variables for each of the 41 cities. Appendix A gives the summary statistics for these variables, including the mean, standard deviation and percentage of the values missing for each variable. We use this section to address particular items of note pertaining to the data set. First, we define our variable set. We use Table 3 and Table 4 to delineate each variable by name and type. The first column in Tables 3 and 4 indicates the variable. This name matches the name of the variable in the summary statistics from Appendix A. The second column of the tables shows the variable name as it's reflected in our actual data set. In the third column, we identify whether each variable is numeric (continuous), integer, binary {0, 1}, or categorical. And, we provide a brief description of each variable in the fourth column.

The degree to which we have missing data is an important component of our study and varies by variable from 0% to approximately 30%. However, by city, the percentage varies from 0% to only approximately 25%. This suggests that the cities tend to have values for the majority of variables, but we also have particular variables that are missing for several cities from our sample. As shown in Appendix A, education has the most missing values, followed by poverty rate. The percentage of the population with a bachelor's degree or higher (Bachelor's Degree [%]) is missing approximately 32% of its values, and a high school diploma or higher (High School Diploma [%]) is missing approximately 27% of its values. Poverty rate contains slight fewer missing values with a total of 20% of its values missing.

We briefly describe the variability in data missing by city and direct the reader to Appendix A for a complete description of missing data. In total, our data set contains 104 missing values, which amounts to approximately 7.6% of the values in the data set. This missing data is concentrated in foreign cities but varies widely in terms of which missing values correspond to a particular city, with the exception of China and India. Within both countries, the data available was consistent internally, which we attribute to the data coming from the same sources. For example, the Chinese and Indian cities have data for between 85–88% of the variables, compared to 98–100% for the U.S. cities. In contrast, cities in Africa, the remainder of Asia, and South America do not exhibit consistency in terms of data availability. Notably, the rest of the developing world has data for between 75–94% of the variables, thereby showing much greater variability in available data.

Table 3. Description of Variables Used to Conduct Clustering Analysis (Pt. 1).

Variable	Variable Name	Variable Type	Description
Electricity Generation (MW)	MegaWatts	Numeric	Total electricity installed capacity in megawatts (MW)
Ground Water Sources	GroundWater	Integer	Number of ground water aquifers
Reservoir/River/Lake Sources	Reservoirs.Rivers	Integer	Total number of lakes, rivers, and reservoirs feeding water delivery
Desalination Plants	Desalination	Integer	Number of desalination plants
Daily Water Dist. (million gal.)	WaterDist	Numeric	Daily amount of water distributed to the municipality in hundred millions of gallons
Int’l Airports	Intl.Airports	Integer	Number of international airports in the metropolitan area
Regional Airports	Reg.Airports	Integer	Number of regional airports in the metropolitan area
Seaport	Seaport	Binary	Whether the city has a seaport
Seaport (million TEUs)	TEUs	Numeric	Annual throughput in millions of twenty-foot container equivalent units (TEU)
Seaport (million tons)	Tons	Numeric	Annual throughput in millions of metric tons
Road Network (miles)	Road.Network	Numeric	Number of miles of road network
Rail Network (miles)	Rail.Network	Numeric	Number of miles rail network
Telecomm Access (%)	Telecomm.Access	Numeric	Percentage population with telecommunications access (mobile or landline)
Subway	Subway	Binary	Whether the city has a subway
Sanitation Access (%)	Sanitation	Numeric	Percentage of population with access to improved sanitation
Hospitals	Hospitals	Integer	Number of hospitals in metropolitan area
Police Stations	Police	Integer	Number of police stations in metropolitan area

Table 4. Description of Variables Used to Conduct Clustering Analysis (Pt. 2).

Variable	Variable Name	Variable Type	Description
Fire Stations	Fire	Integer	Number of fire stations in the metropolitan area
Military Bases	Military	Integer	Number of military bases in metropolitan area
State Capital	State.Capital	Binary	Whether the city is a state/ provincial capital
National Capital	Natl.Capital	Binary	Whether the city is a national capital
GDP (billion USD)	GDP	Numeric	Annual GDP
High School Diploma (%)	High.School	Numeric	Percentage of population with a high school diploma or equivalent
Bachelor’s Degree (%)	Bachelors	Numeric	Percentage of population with a bachelor’s degree
Literacy Rate (%)	Literacy.Rate	Numeric	Percentage of population deemed literate by local standards
Average Income	Avg.Income	Numeric	Average annual household income in U.S. dollars
Poverty Rate (%)	Poverty.Rate	Numeric	Percentage of population living below the local poverty line
Average Household Size	Household.Size	Numeric	Average number of people living in each household
Pop Under 18–19 Yrs Old (%)	Under.18	Numeric	Percentage of population under 18 years old or 19 years old, whichever is available
Pop Over 65 Yrs Old (%)	Over.65	Numeric	Percentage of population 65 years old and older
Internet Access (%)	Internet.Access	Numeric	Percentage of population with Internet access in the metropolitan area
Religion	Primary.Religion	Categorical	Religious affiliation of the largest percentage of the population. This variable has six levels. Christian, Islam, Chinese religion or atheist, Hinduism, Buddhism, and Vietnamese Religion or atheist
Pop Density (Per sq. mile)	Pop.Density	Numeric	Number of people per square mile in the metropolitan area.

Across the full set of variables, we also notice that variables are measured on very different scales and some exhibit quite a bit of variability (see Appendix A for standard deviations by variable). In Figure 4, we show side-by-side comparisons of the boxplots for each of the 32 numeric, integer, and binary variables in our data. We standardize each variable to a mean equal to zero and a standard deviation equal to one so that they are on the same scale.

Figure 4. Comparison of Boxplots for Each Variable in Study, Standardized to a Mean=0 and Standard Deviation = 1.

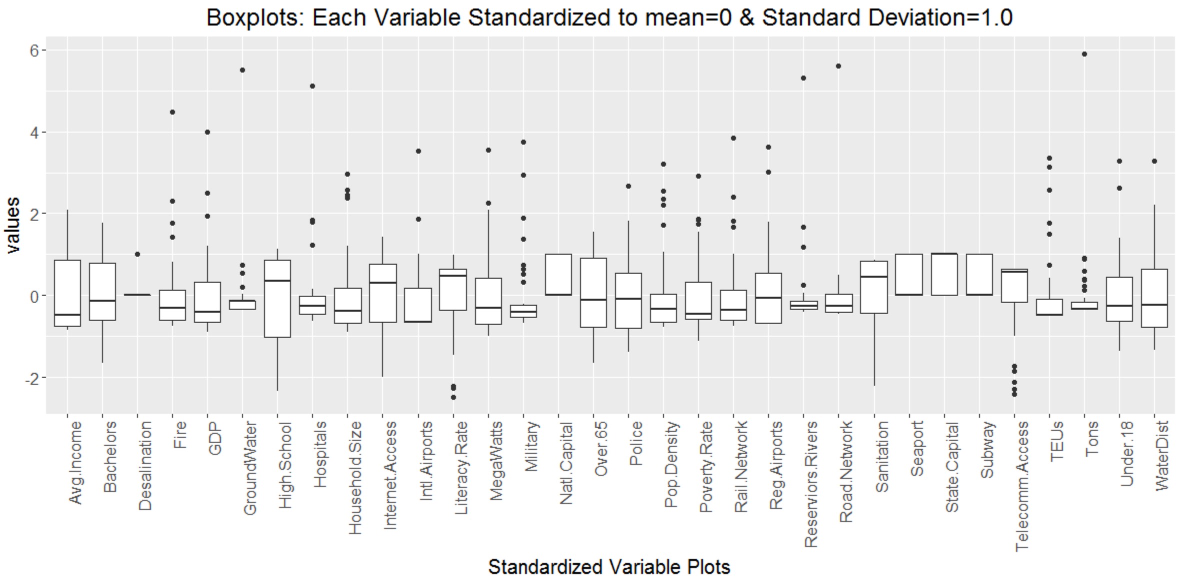


Figure 4 shows the distribution of each numeric (continuous), integer, and binary {0, 1} variable in our data set. Each variable is standardized to a mean of zero and standard deviation of one. Clearly, the majority of the data is right skewed with large outliers. However, we also see that literacy rate and telecommunications access are left skewed with small outliers. Sources used for compilation can be found in Appendix C, and plotted using methods from Wickham (2009).

We can see that the majority of the variables shown in Figure 4 are right skewed. In particular, we note some of the variables that have extremely large outliers with a large separation from the next largest value. New York City is an extreme outlier in the number of fire stations (673) supporting its MSA, the Washington, DC, MSA is a large outlier in ground water sources (33), Chongqing is the largest outlier in number of hospitals (1502) and length of their road network (87,364 miles), Johannesburg has the most rivers and reservoirs (87) that feed its water distribution system, and Shanghai has the largest tonnage of annual throughput (498 million tons) at its seaport. Each of these values are over two standard deviations away from the next largest value. In contrast, we also see that the data have some variables whose distributions are left skewed. Dhaka has the lowest literacy rate (43.6%), and Chongqing has the lowest access to telecommunications (14.1%). But, these outliers are within one standard deviation of the nearest values, Chicago and Kabul, respectively. The shapes of variable distributions and the extreme outliers impact the techniques we can apply to impute missing data and cluster our cities.

In addition to the summary information in Section C, we also include our data sources and specific calculation methods for certain proxy variables in Appendix B and Appendix C. Specifically, we dedicate Appendix B to addressing how we approximate variables where data is available but not in the form we seek. This includes accounting for values for variables that are only available at that state or provincial level, or variables that do not directly measure a particular feature we seek to analyze but serve as viable proxies. For example, in some cities installed electricity capacity is not always available, so we outline how we incorporate proxies such as electricity consumption or electricity generation. In Appendix C, we identify, by country, the sources for each of the 33 variables in our study.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. METHODOLOGY

In this chapter, we detail the methods used to form the clusters, how we handle the missing values in the data, and the techniques required to produce an ensemble of clusters and “average” clustering, based on the randomized missing data imputations.

A. DISTANCE CALCULATION

Prior to producing data clusters using current techniques, we first identify the method we will use for calculating the distances between observations. Several methods are available that produce generally good results, but the particular method to use often depends on the nature of the data. Our data contains three features that limit our options. Specifically, the numeric variables are measured on very different scales which can skew results if they are not handled properly. Additionally, approximately 8% of our data are missing. Although we intend to impute the missing values, we also want a distance measure that can handle missing values for our initial exploration and determining the appropriate number of clusters. Finally, our data also contains a categorical variable with six levels. This challenge could be mitigated by using six different binary variables to account for each level, but we seek a distance measure that automatically accommodates categorical as well as numeric variables without requiring extra data manipulation.

Euclidean and Manhattan distances are common choices, but they are limited to handling numeric variables and do not perform as well when the data contain outliers, like ours. We elect to use Gower’s (1971) method to calculate our distances because it is capable of handling missing values, mixed-type quantitative variables, and categorical variables and internally standardizes data. If we have a n x p matrix of data where n is the number of observations and p is the number of variables, Gower’s coefficient uses the absolute difference between two observations i, j, on a variable k, divided by the range of the k^{th} variable. This computes the coefficient for quantitative variables (Gower, 1971). In the case of categorical variables, Gower (1971) computes differences by assigning a 1 if observation i and j match for a given variable and 0 if they do not. These calculations also manage missing values. In general, Gower (1971) ensures that missing values are not directly included in computing the distance between observations *i* and *j* by assigning a weight of 1 if the variable is present in both observations and 0 if one or both are missing. Finally, he calculates the total distance coefficient using the average distance coefficient between observations across all variables. We do not show the mathematical equations that produce the Gower's distance in detail, so we recommend readers see Gower (1971) for further explanation.

We believe Gower's method provides us with a reasonable technique for calculating the differences among our megacities. In order to implement the calculations for Gower's distances, we use the DAISY function from the cluster package (Maechler, et al., 2016) in the R programming language (R Development Core Team, 2008). DAISY provides the capability to implement three major distance calculations including Euclidean, Manhattan, and Gower distances by taking the data set as an input and producing an output of the resulting *m x n* matrix of all pairs of inter-point distances. We can then use this output as input for our clustering algorithms.

B. CLUSTERING ALGORITHM

We rely on a simple partitioning method to cluster the 41 megacities into distinct group. We will not address all possible partitioning methods, but we mention the most common of these techniques including PAM, Clustering Large Applications (CLARA), and K-Means. The K-Means algorithm uses Euclidean distances between observations and produces clusters based on each observation's nearest neighbors such that the within cluster distance is minimized (MacQueen, 1967). While similar to K-Means, PAM and CLARA seek to identify central observations called "medoids" that serve as representative observations for a cluster (Kaufman & Rousseuw, 1990). PAM and CLARA then pair each observation to its nearest medoid in order to produce the clusters (Kaufman & Rousseuw, 1990). We differentiate PAM and CLARA by noting that CLARA is for use in large datasets and uses sampling techniques to identify medoids rather than the full set of observations, as used in PAM. For the purposes of our study, we elect to use the PAM algorithm from Kaufman and Rousseuw (1990). Among its desirable traits, PAM allows distances between observations to be defined by the user whereas K-Means requires all variables to be numeric and distances Euclidean. Thus, PAM, allows for data with categorical variables.

C. NUMBER OF CLUSTERS

To use the PAM algorithm, we must first identify the number of clusters to use. An initial perusal of the work done in United States Army (2014) and Sapol (2016) would lead us to select three clusters or six clusters, respectively, as our baseline. However, this assumes that the number of categories in those studies were the best possible options. To choose the appropriate number of clusters, we use an internal validation measure known as the silhouette width, which is calculated for individual observations and then averaged over individuals for each cluster as well as the whole data set. For the mathematical definition of silhouette width, see Kaufman and Rousseuw (1990). On the individual observation level, it identifies how distinct a given observation is from its nearest neighboring cluster relative to its actual cluster. When averaged, it represents a signal of how distinct each cluster is from other clusters and the degree to which the data has an underlying clustering structure. Silhouette width ranges from -1 to 1 where values closer to 1 represent the most distinct clustering.

To be clear, let us first define two separate clusters as C or D. When the silhouette of an observation i from cluster C is close to 1, the distance to observations within its clusters is much smaller than its distance to its nearest neighboring cluster, D (Kaufman & Rousseuw, 1990). As the silhouette approaches 0, we see that observation i becomes neutral between cluster C and D. Finally, as the silhouette approaches -1, we see the worst case scenario (Kaufman & Rousseuw, 1990). In this case, observation i has smaller relative distance to its nearest neighbor cluster D than its actual cluster, so we expect that misclassification likely occurred in placing i in cluster C rather than D (Kaufman & Rousseuw, 1990).

The average silhouette width, aggregated at the full data set level, can then be seen as a measure of the overall performance of the algorithm for a given number of clusters. In order to establish a baseline, we produce the average silhouette width for each number of clusters of the data ranging from $c = 2$ to $c = 10$, where c is the number of clusters. We select $c = 10$ as our upper bound because we only have 41 observations, and we expect more than 10 clusters will limit the insights that we can glean from the data. At the same time, we also do not choose a very small number of clusters simply because it produces the maximum average silhouette width. If a larger number of clusters produces a reasonable relative average silhouette width, we look to the larger number for our selection because it produces more diversity in the groupings.

D. MISSING VALUE IMPUTATION

Missing values are a significant consideration for our data set, so we must handle them with care. In order to impute the missing information, we elect to use the K-Nearest Neighbor (K-NN) quantile sampling method adapted in Kowarik and Templ (2016). Because each variable has approximately 70% or more of its data available, we believe the information available is sufficient for producing good imputations of the missing values by using the observations to which they are most similar. We choose this method over other options because we believe the others do not fit the structure of our data as well, and K-NN quantile sampling allows us to assess sensitivity of the clustering results to the missing values.

The simplest form of data imputation involves simply using the mean, median, or in the case of categorical variables, the mode, in place of the missing values. But, this method does not capture the uncertainty or variability of the missing information or dependence among variables. To account for variability, missing values are often randomly imputed numerous times. And, they are sometimes generated from parametric distributions, but these typically require assumptions regarding the marginal distributions and independence among variables. For numeric variables, random normal variates can be generated with mean equal to the sample mean of the j^{th} variable and standard deviation equal to the sample standard deviation of the variable j. Given we have categorical variables and the remaining variables are right or left skewed, normal distributions do not adequately capture the natural variation in our data. Other distributions like the triangular may be used in place of the normal distribution, but this approach also requires us to assume that the marginal distributions are independent. In experimentation, we do impute numeric missing values using the triangular distribution where the minimum and maximum values are taken to be the first and fourth quantiles for each variable, and the mode is taken to be the median. But, our education variables (high school and bachelor's degree) and our age distribution variables (under 18 and over 65) are dependent. We require a method for final analysis that captures the dependence.

Another common method for handling missing data uses bootstrapping techniques. In a basic bootstrap, we sample values from those observed in column j and replace those missing in variable j with the randomly selected values. This provides us with variability and maintains the marginal distributions of the variables, but we risk including the values associated with extreme outliers in locations where we know that we cannot reasonably expect them to exist. For example, the data available for Chongqing included a road network of over 87,000 miles (China Data Online, 2017), the next closest city using Gower distance was Delhi with a road network of almost 14,000 miles (Indiastat, 2017), and the mean road network length for the full data set was only approximately 6,500 miles. With five of the values for the road network variable missing, we find the risk of Chongqing's road network being used as the imputed value for other cities to be unreasonable when it can be avoided.

Van Buuren (2012), also provides several techniques that can be used to impute missing data. Each of them has strengths and weaknesses and can be used for binary, continuous, or categorical variables. His Multiple Imputation by Chained Equations (MICE) algorithm samples from the observed data and imputes the missing data on a variable-by-variable basis and repeats the imputation multiple times in order to produce multiple datasets. MICE can perform a combination of univariate imputation techniques, which allow the user to control the method applied to each individual column, including using the mean, bootstrapping, linear regression, and stochastic regression among others for data transformation that do not apply to here (Van Buuren, 2012). Unfortunately, his work generally is not intended to perform particularly well in the presence of small data sets with data missing not at random (MNAR), as is the case in this study.

As a result of these challenges, we take on the approach of random sampling from the data set using characteristics inherent to the data. The K-NN imputation algorithm calculates the distance between observations using an extension of Gower's method (Kowarik & Templ, 2016). Then, using the calculated distances, it identifies the k observations to which it is closest. One major benefit is that this method allows the user to select the value of k. Once the k-nearest neighbors are identified, the K-NN algorithm uses the statistical properties of those nearest neighbors to calculate an imputed value for the observation with missing information (Kowarik & Templ, 2016). For example, say we have a missing value for Hyderabad in poverty rate, and its $k = 4$ nearest neighbors are Delhi, Mumbai, Kolkata, and Bangalore. As a default, the K-NN algorithm will calculate the median poverty rate for Delhi, Mumbai, Kolkata, and Bangalore, and then it will impute that number into the missing value for Hyderabad. We call this K-NN median imputation. While this gives us some additional information, it is not stochastic in the sense that the imputed value for the nearest neighbors will not change. Fortunately, the K-NN algorithm also allows the user to create their own aggregation method to calculate the imputed value, which significantly increases the flexibility of the algorithm (Kowarik & Templ, 2016). The aggregation method simply defines how the K-NN algorithm will use the data from the nearest neighbors to calculate the missing value.

As we discuss earlier in Chapter IV, most of our data is right or left skewed. Recall from Figure 4 in Chapter III showing the side-by-side boxplots that poverty rate is right skewed due to a few high values like the 62% poverty rate found for Lagos (Open Data for Africa, 2017a). Extreme values like this one shift the mean poverty rate to 18.65%, while the median value is 11.98%. Therefore, by using the K-NN algorithm, we are much better positioned to capture a likely value for the missing data because we use similar observations. We add randomness for numeric variables by randomly generating an imputed value from a uniform distribution between the minimum and maximum value of the k-nearest neighbors. A discrete uniform is used for integer variables and a continuous uniform is used for others. Hence, we are not limited to only producing values that appear in the data set. This allows any value within that range to be imputed for continuous variables and any integer to be imputed for discrete variables. We do not implement a technique for categorical variables because our binary indicators in our data set are treated as integers, and the categorical variable for religion does not have missing values. In what follows, we call this type of imputation K-NN quantile imputation because we use the minimum and maximum – the smallest and largest empirical quartiles from the k-nearest neighbors.

E. CLUSTER ENSEMBLES

Using the terminology from Hornik (2005), a “clustering” is a partition of the data set that has been divided into *c* groups. In contrast, clusters are individual groups labeled from 1 to *c* to identify the group within the clustering. Hornik (2005) outlines the methodology for producing multiple clusterings for a given set of observation, for example, re-sampling the data. Sets of clusterings are known as ensembles. Hornik (2005) also defines distances between clusterings that allow us to identify the levels to which the different clusterings agree or disagree. If clusterings *A* and *B* agree, we expect that the observations that cluster together in clustering *A* will also cluster together in clustering *B*. This is particularly beneficial for our study where we are identifying the effects of changes in imputed values.

We use the R packages CLUE (Hornik 2005) to take an ensemble of clusterings and calculate the distance between pairs in the ensemble. CLUE can also produce what Hornik (2005) refers to as the consensus cluster. We recommend readers review Hornik (2005) for further review of his methods for ensembles of clusterings. In essence, the consensus cluster reveals the cluster to which each observation belongs, “averaged” over the clusterings in the ensemble. With clusterings formed by randomly imputing large amounts of missing data, this gives us a sense of the typical clustering. We can also compare this consensus clustering to results obtained from clustering with missing values in order to observe any meaningful changes.

Hornik (2005) constructs the consensus clustering through both *hard* and *soft* clustering. In hard clustering, each observation is identified with only one cluster. Soft clustering assigns a weight to each cluster for each observation. If there are *c* clusters in the result, an observation will have a weight of between 0 and 1 for each of the *c* clusters, which sum to 1. In cases of soft clustering where an observation has a weight close to or equal to 1, we can say that the observation generally belongs with the other observations in that cluster. And, as the weight decreases and is spread across clusters, we can say that its assignment may be unclear or indifferent between clusters on average. This captures the uncertainties that we seek in identifying how sensitive our clusters are to changes in the missing data.

For our work, this capability allows us to impute missing values several times and then use the resulting data sets with missing values to produce a clustering for each one separately. Specifically, we use the K-NN imputation technique to generate 5,000 different data sets of our 41 observations. Using Gower distances with PAM clustering, we compute 5,000 sample clusterings corresponding to our 5,000 data sets. With these, we then use the cluster ensemble algorithm from the CLUE package in order to produce an ensemble of clusters, from which we calculate a clustering distance matrix. The clustering distance matrix gives us the distances between each pair of clusterings in the ensemble and enables us to generate both the hard and soft consensus clusterings. We use classical multidimensional scaling to then map the 5,000 clusterings to a two dimensional space, such that the Euclidean distance between points in the two dimensional space is approximately equal to their respective clusterings' distances (Gower, 1966). This allows us to observe the variability in the clusterings because we can visualize each clusterings' location relative to all others in a standard Cartesian coordinate plot. This also permits us to see how similar our exploratory clusterings are to clusterings in the ensemble and to each other. Specifically, we compare the ensemble to the baseline, the consensus, the K-NN median, and to a realization of a clustering using missing data imputed from marginal triangular distributions.

V. RESULTS AND ANALYSIS

In this chapter, we build clusterings using the techniques outlined in Chapter IV to identify the underlying structure of the cities in our data set. First, we determine the number of clusters to use by analyzing the base data using Gower's distance. We then discuss in detail the clustering results for this number of clusters and no imputation of missing values. In order to capture the uncertainty of the missing values, we then use K-NN quantile sampling to impute the missing values and form an ensemble of 5,000 clusterings, from which we compute consensus clustering. Finally, we draw comparisons between the clusterings and provide a final discussion comparing our results to those found in previous work.

A. SELECTING THE NUMBER OF CLUSTERS

When we calculate the average silhouette width for $c = 2$ to $c = 10$, we find that $c = 4$ produced the best average silhouette width using Gower's distances and PAM for clustering. It produced a value of 0.238, which suggests there may be very little or no underlying structure in the data according to Kaufman and Rousseuw (1990). However, they also concede that this is a subjective guide based on their experience and may not apply a particular study. Hence, we recognize this as a challenge but do not find it particularly concerning for this study. In comparison, the next largest value in the average silhouette width plot proved to be $c = 5$ with a value of 0.237. While this is slightly lower than the average silhouette width for $c = 4$, by using $c = 5$, we have an additional benefit of greater diversity in the clusters and the potential for more useful results. Because we seek to understand which cities are similar or different in interesting ways, a more diverse collection of clusters likely reduces the number of cities in each cluster. Generally, we expect this provides us greater insight into which cities are most similar. As a result, we elect to use $c = 5$ as the appropriate number of clusters for the remainder of the study.

B. CLUSTERING WITH MISSING VALUES

We use the PAM algorithm and Gower distance calculations to produce five clusters. We accomplish this without imputing missing data in order to establish a baseline clustering and allow us to see what changes occur when we incorporate imputed data. The results of this clustering can be seen in Table 5. We indicate the number of cities in each cluster in parentheses next to the cluster number. Additionally, as discussed in Chapter IV, PAM uses medoids as the representative cities for each cluster. We indicate these representative cities in each cluster with an M in parentheses next to the name of the city. These medoids provide us with a barometer for what we can expect in the features of variables for other cities in the cluster.

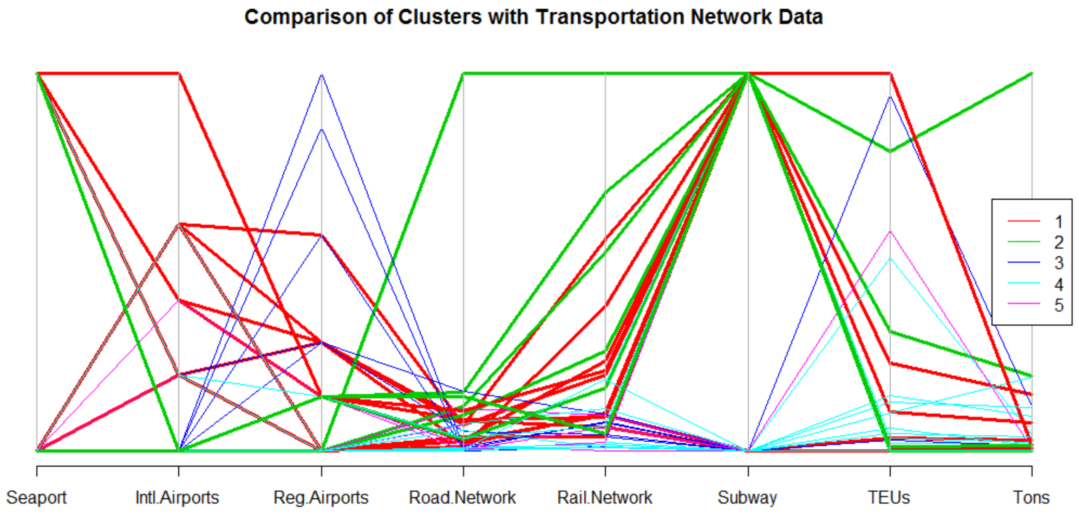
Table 5. Clusters Using Missing Values and c = 5

Cluster 1 (10)	Cluster 2 (6)	Cluster 3 (6)	Cluster 4 (12)	Cluster 5 (7)
New York Los Angeles Chicago Washington, DC Dallas-FW San Francisco Philadelphia (M) Boston Toronto London	Beijing Tianjin (M) Shanghai Chongqing São Paulo Rio de Janeiro	Delhi Mumbai Kolkata Bangalore (M) Hyderabad Johannesburg	Lagos Nairobi Baghdad Jakarta Bangkok Manila Mexico City Moscow Istanbul Buenos Aires Lima (M) Ho Chi Minh City	Cairo Kinshasa Karachi Al-Riyadh Tehran Kabul Dhaka (M)

The table shows which cities fall into each one of the five clusters. In parentheses, we show the total number of cities in each cluster next to each cluster label. And, the “M” next to one city in each cluster represents that city being a “medoid” or the representative city for that cluster.

Of note, the western cities tended to cluster together despite their differences. These cities vary in non-trivial ways with regard to their infrastructure and economic power, and yet, they are still more similar within that cluster than they are to other cities. Additionally, Chinese cities and Indian cities dominate clusters two and three, respectively. In contrast the Middle Eastern, African, South American, and Southeast Asian cities spread across multiple clusters, which illustrates geographical proximity may not necessarily dictate similarities or differences among cities. South American cities are split between clusters two and four, and African cities are split between clusters three, four, and five. When we examine the clusters using parallel coordinate plots (Venables & Ripley, 2002) we are able to see if particular variables drive cities into one cluster or another. Figure 5 uses the transportation network variables including the presence of a seaport, number of airports, the road network, and rail networks to identify key differences in the data for each cluster. Each number in the legend corresponds to its related cluster number from Table 5. For example, the number 4 line in Figure 5 corresponds to the cities located in cluster 4. All variables in Figure 5 are transformed to be between 0 and 1. Missing values are not plotted and can be seen as discontinuities (or breaks) in a city's line.

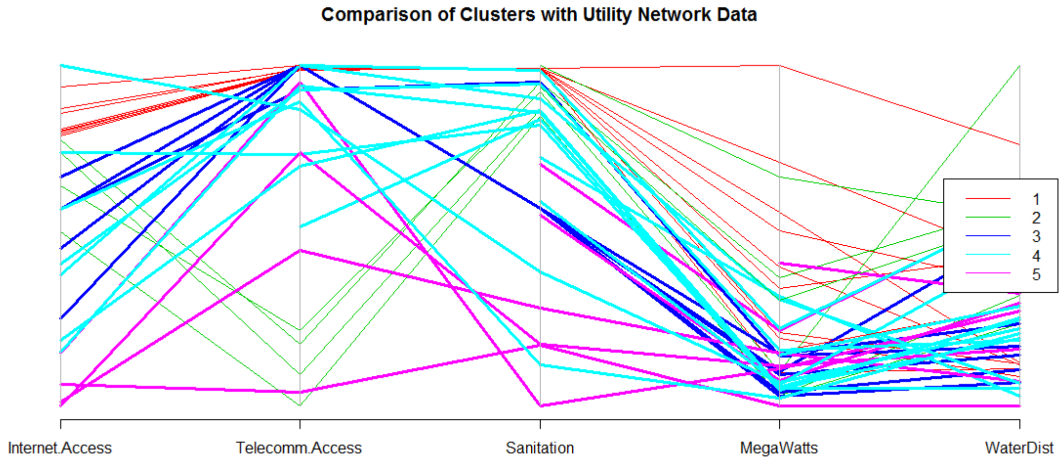
Figure 5. Comparison of Base Clusters Using Transportation Network Data



In Figure 5, the legend indicates the cluster number associated with each color and the vertical axis indicates the range of possible values for each variable, where each line represents a different observation. We note that Subway and Seaport are binary, which shows the clusters splitting between those with that type of infrastructure and those without. The Indian cities (in blue) cluster around a large number of regional airports, and the Middle Eastern, African, and Southeast Asian cities in clusters 4 and 5 cluster around no subways and the smallest road networks and rail networks. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

When we examine the parallel coordinate plots, we see that cities without a subway like Karachi, Tehran, and Kabul clustered together and tend to have less robust road and rail networks. In contrast, cities in the U.S. (cluster 1 in red) tended to have more robust transportation networks and more international airports. Interestingly, the cities in cluster 3, mostly represented by India, tended to have a larger number of regional airports. Throughput through seaports in TEUs (twenty-foot equivalent units) and tons do not seem to clearly delineate clusters to the same degree as some of the other variables, with the exception of a large number of cities from cluster 4 in the middle of the seaport cargo throughput range in TEUs. We also note the large outlier in the road network and rail network variables from cluster 2. As discussed in Chapter IV, Chongqing has extreme values for variables, but they do not seem to exert undue influence over how the cities cluster. Next, we examine the parallel coordinate plots (Venables & Ripley, 2002) for utilities such as Internet, telecommunications, sanitation, electricity, and water using Figure 6.

Figure 6. Comparison of Clusters Using Utility Network Data



In Figure 6, the western cities in cluster 1 dominate the upper bounds in access to utilities. Cities in cluster 5 have low access, and cluster 3 is inconsistent. In some cases the Chinese cities (in green) cluster toward the bottom of access and in others they cluster toward the top, as is the case in telecommunications access and sanitation. Beyond author knowledge of the data set, we can tell they are the Chinese cities the way they move in unison for each variable, much the same way the India or U.S. cities cluster together. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

In Figure 6, we see that the western cities reside primarily at the upper bounds of each variable pertaining to utility access. The U.S., Canada, and United Kingdom all have robust utility networks and broad access to telecommunications, clean water, and sanitation. Cities in China that make up the majority of cluster 2 tend to be inconsistent with where they fall in the ranges of values. In the case of telecommunications and electricity they reside near the bottom, but they have broad access to the Internet and sanitation. The Indian cities from cluster 3 have good telecommunications access, but they tend to have lower values for some of the other variables. In contrast, the variability in cities from clusters 4 and 5 seems to be greater than that of the other 3 clusters. For example, cities from cluster 4 have all ranges of Internet access and sanitation access levels, which shows they are interspersed among other clusters in these categories. This may influence distance calculations and cause some cities in clusters where this occurs to be less dissimilar to neighboring clusters than we desire, thereby reducing the performance of the clustering algorithm.

We examine the plots (Venables & Ripley, 2002) for access to emergency services in Figure 7. This plot compares the clusters for the number of hospitals, police stations, fire stations, military bases, and indicators for state or national capitals.

Figure 7. Comparison of Clusters with Emergency Service Access and Government Data

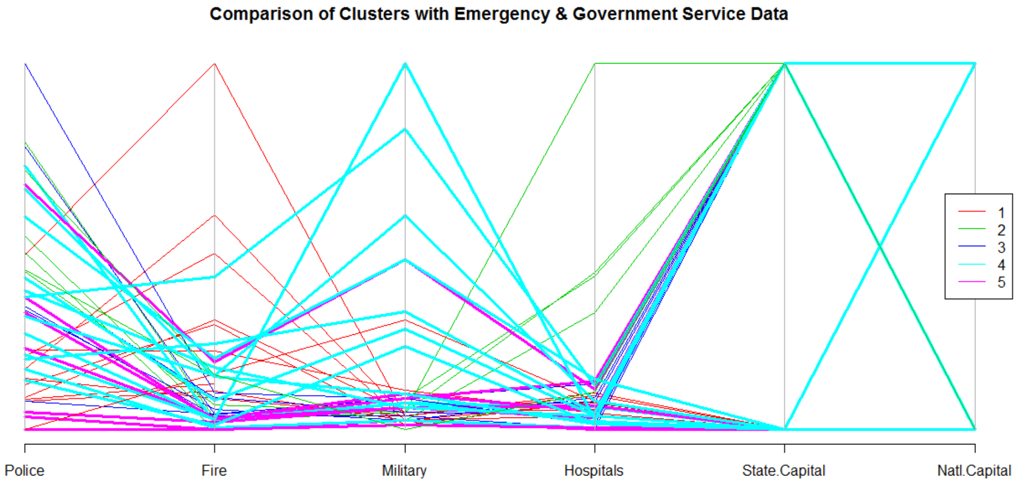


Figure 7 shows that the data for emergency service and government infrastructure do not clearly separate the ranges of data into clusters. But, cities from cluster 1 in the west tend to have the most fire stations and are primarily the only cities that are not state or national capitals, while cities in cluster 4 tend to have the most military bases. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

We see that the data for access to emergency services and government infrastructure do not seem to clearly separate cities into clusters with the exception of a few examples. The western cities from cluster 1 tended to have more fire stations and are generally not state or national capitals. And, cities in cluster 4 tend to have a large number of military bases. While we believe this type of data is important to analyzing the differences among megacities, we recognize the potential weaknesses it poses for our analysis. Improvements in data quality through more accurate sources or different indicators may improve results in future studies. For example, instead of the absolute totals for fire stations, police stations, and hospitals, future work can examine the per capita values. Or, hospital data could focus on the number of beds or operating rooms per 1000 people, and police data may also include violent and non-violent crime rates per 1000 people.

We show the resulting clusters for economic data in Figure 8. The economic data includes the GDP/GRDP, average household income, poverty rate, percent of high school or equivalent graduates, and percent of persons with bachelor's degrees. As discussed in Chapter III, these variables provide an indication of the overall economic health of a particular city.

Figure 8. Comparison of Clusters with Economic Data

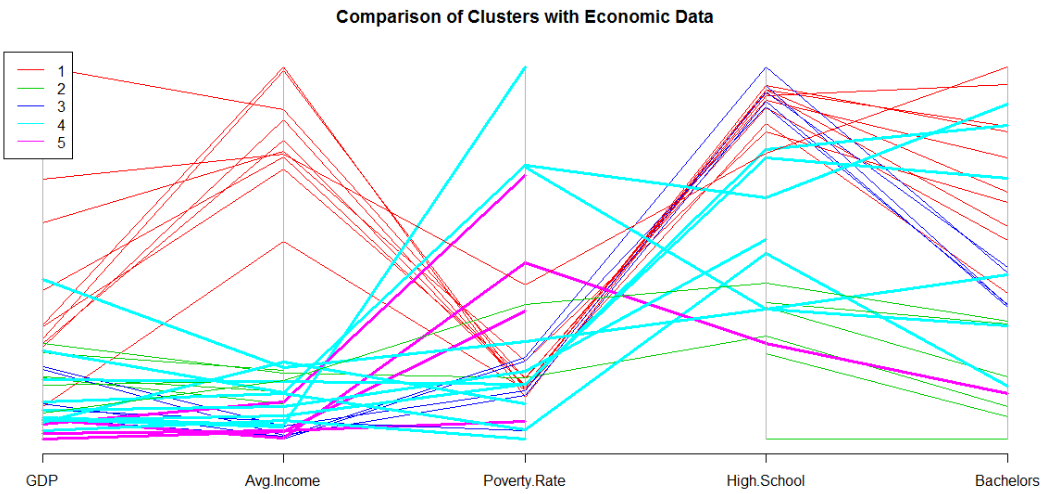


Figure 8 shows that the cities in cluster 1 again diverge from the remaining urban areas. With large average incomes and GDP, high percentages of educated people, and low poverty rates, these cities cluster well. The cities in cluster 5 generally have the lowest incomes and GDP and relatively higher poverty rates. We also note that the majority of cluster 5 has missing data for education rates, annotated by breaks in their lines. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

The clusters illustrate which cities cluster around strong economies, which cities are in developing economies, and which cities generally have the weakest economies. The economic data reveals clusters in the data more effectively than government and emergency services, but we continue to see large groups of observations at the upper or lower ends of the spectrum. Notably, western and Indian cities cluster together at high levels of high school education, but the western cities separate to higher levels of bachelor's education and above. All cities in clusters 2 thru 5 have lower average incomes and GDP/GRDP than the cities in cluster 1. However, within clusters 2 thru 5, the cities are interspersed and do not distinguish themselves into groups effectively. Hence, these variables may also negatively influence the average dissimilarity between clusters. Interestingly, the cities in cluster 5 tend to have missing values in their education variables, which may impact the ability of additional information to shift the cluster locations of those cities including the cities in the Middle East, Kinshasa, and Dhaka.

In the results for demographic variables, we show the age distribution, average household size, population density, and literacy rate. Figure 9 shows how these variables influence the cluster of our data with the missing values.

Figure 9. Comparison of Clusters with Demographic Data

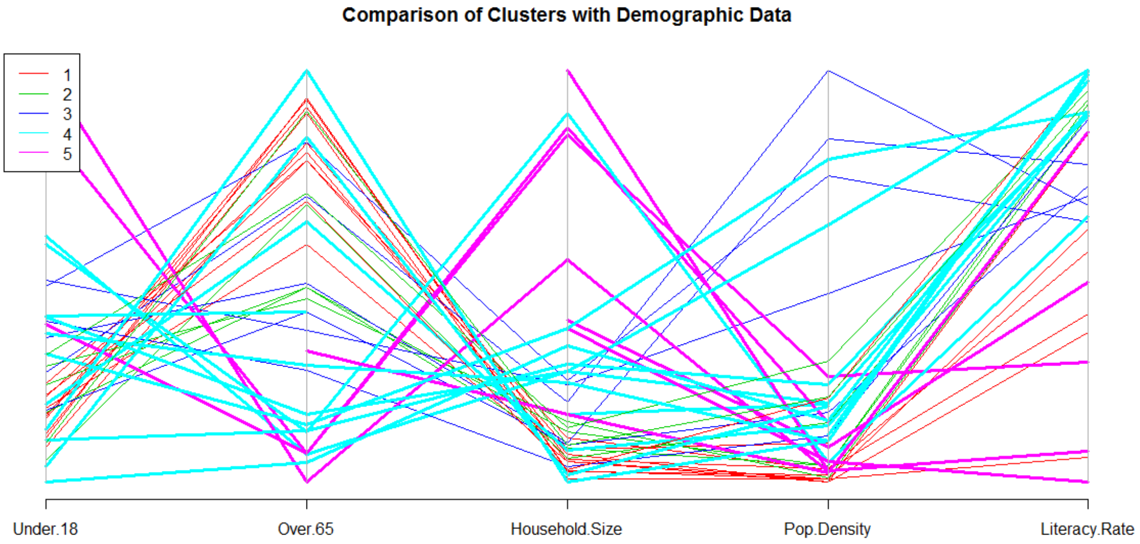


Figure 9 shows the dispersion of clusters among various levels in the demographic factors. Breaks in the lines for observations indicate that the variable is missing for a given observation. Notably, Cluster 5 cities tend to have missing information in their demography, but the information available seems to correspond to larger household sizes, less people over 65 years old, and more young people. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

We see in Figure 9 that clear delineations between clusters are also difficult when comparing demographic data. We notice that the U.S. cities surprisingly reside toward the bottom in literacy despite their strong education rates displayed in Figure 8. This is an interesting dynamic and may point to differences in how different cities and countries define literacy standards. The cities in cluster 5 have large amounts of missing information in their age distributions as seen by the breaks in their lines going from variable to variable. However, the information available suggests that they are somewhat similar to cities in cluster 4 in the sense of having relatively high percentages of young people and relatively low percentages of older people 65 and up. This generally matches with conventional wisdom and data for their countries. According to the Knoema (2017), only 2.5% of Afghanistan's and 5.0% of Bangladesh's populations have ages 65 or above.

As we discuss earlier in Chapter IV, the observations we expect to be most at risk of transitioning between clusters in the presence of imputed values are those with low silhouette widths and observations with more missing data than other observations within their cluster. Observations with small silhouette widths (≤ 0) are more similar to their nearest neighboring cluster than their actual cluster, which, when coupled with more missing data, clearly creates more opportunity for an observation to move between clusters. This also applies to observations in the neighboring cluster. The addition of imputed data will inevitably change the Gower's coefficient between two observations, particularly those with more missing data. And, as a result, the changes may drive them closer together or farther apart.

To further illustrate this, we also note inter-cluster average silhouette widths and some key observations with particularly low silhouette widths that may be prone to shifting with imputed data. The average silhouette widths for each cluster are displayed in Table 6. Recall, the average silhouette width associated with a cluster is the sum of all the observation silhouettes in the cluster divided by the number of observations in the cluster. While this generally signals how distinct a cluster is relative to all others, we note that there can certainly be observations within a cluster that are well above or well below the average.

Table 6. Average Silhouette Width for Each Cluster with Missing Data and c=5

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
0.37	0.33	0.41	0.03	0.18

Table 6 shows that cluster 3 with Indian cities is the most distinct from the others on average, with the next being cluster 1 with the western cities. Cluster 4 contains mostly African and Southeast Asian cities, and this represents a likely choice for movement due to changes in the data, followed by cluster 5. Recall, silhouettes near 0 signify indifference between an observation's cluster and its nearest neighboring cluster.

Clusters 1, 2, and 3, containing the western cities, Chinese cities, and Indian cities, respectively, are clearly the most distinct. Then, we notice a large drop in the average silhouette widths for clusters 4 and 5. This is important to note because, these observations tend to have the most missing values in addition to not being significantly distinct from their nearest neighbors. Specific to cluster 4, Lagos, Nairobi, Istanbul, Baghdad, and Moscow each have negative silhouettes relative to their nearest neighboring cluster. Our results also show their next nearest neighbor being clusters 5, 5, 3, 5, and 1, respectively. For example, these results indicate that Lagos is actually more similar to observations in cluster 5 than cluster 4 and may have been misclassified. Surprisingly, these results indicate that Moscow is more similar to the western cities in cluster 1 than its current cluster (4). Hence, we may expect them to be "first movers" and transition to their nearest neighboring cluster in the presence of changing data. In contrast, all cities in clusters 1, 2, and 3 have relatively higher silhouette widths. Hence, while some observations may change due to the stochastic nature of the data imputation, we do not expect significant movements from cities currently in these clusters.

C. CLUSTERING WITH DATA IMPUTATION

We begin clustering using PAM with Gower distances based on data imputation using K-NN median imputed values with $k = 5$ for $c = 2, \dots, 10$ to see if the data imputation produces dramatic changes in the number of clusters. Our results indicate that the data imputation does not change the results in a significant way. In fact, $c = 5$ clusters produces a larger average silhouette width than $c = 4$ clusters with totals of 0.271 and 0.259, respectively.

The resulting clusters can be seen in Table 7. They illustrate that imputation of the missing data can influence the clusters to which some of the cities belong. And, the cities most impacted are the cities with more missing data. Recall, K-NN median imputation is not stochastic in nature, so we do not require multiple samples. We show this result as a building block to the ensemble of clusterings we will generate in Section D.

Table 7. Clusters Using K-NN Median Imputed Data and c=5

Cluster 1 (11)	Cluster 2 (5)	Cluster 3 (6)	Cluster 4 (8)	Cluster 5 (10)
New York	Beijing	Delhi	Lagos	Cairo
Los Angeles	Tianjin (M)	Mumbai	Jakarta	Kinshasa
Chicago	Shanghai	Kolkata	Bangkok	Al-Riyadh
Washington, DC	Chongqing	Bangalore (M)	Manila	Kabul
Dallas-FW	São Paulo	Hyderabad	Buenos Aires	Karachi
San Francisco	Rio de Janeiro	Johannesburg	Lima (M)	Tehran
Philadelphia (M)			Ho Chi Minh City	Dhaka
Boston			Istanbul	Nairobi[4]
Toronto				Baghdad[4] (M)
London				Mexico City[4]
Moscow[4]				

Table 7 illustrates how the clusters shift in the presence of one run of imputed data. The cities highlighted in yellow changed, and their corresponding number in brackets indicates the cluster number from the results in Table 5. We note improved balance among the number of clusters in each category and changes in the medoids. Baghdad changed clusters and became a medoid.

We see in Table 7 that four cities change clusters when we add the imputed data. We also find it interesting that Baghdad shifted from cluster 4 to cluster 5 and simultaneously became the medoid for cluster 5. Clearly, the Chinese, U.S., Brazilian, and Indian cities remain consistent in their clusterings. We find this interesting because U.S. cities accounted for the only cities within those groups that contain no missing values.

As discussed at the end of Section B in Chapter V, we obtain some movement from cities with low silhouettes. Notably, Nairobi, Moscow, and Baghdad each move to their nearest neighboring cluster. However, Lagos remains in cluster 4 despite having a negative silhouette in the original clustering. These results improve the average silhouette widths for each cluster when we use the median of the K-NN. And, this makes sense theoretically because we are drawing observations nearer to the observations to which they are already closest. Table 8 shows the average silhouette widths when we incorporate the K-NN algorithm with the median, compared to the base results.

Table 8. Comparison of Base Data with Missing Values and K-NN Median Imputed Average Silhouette Widths

Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Missing Data	0.37	0.33	0.41	0.03	0.18
K-NN Median	0.35	0.36	0.38	0.12	0.18
Percent Change (%)	-5.4	9.1	-7.3	300.00	0.0

Table 8 shows that using K-NN median imputation significantly improves the average silhouette width for cluster 4, which makes sense because approximate 34% of the missing values in our data set come from cluster 4. Therefore, when imputing data to nearest neighbors, we can expect they will become similar in a more significant way than other clusters.

From Table 8, we note that the average silhouette width of cluster 4 increases by approximately 300%. This indicates that the data imputation improved the similarities among the cities in cluster 4 by transitioning to some other clusters and making those within cluster 4 more similar. Given that approximately 34% of the missing values in our base data set come from cluster 4, it is reasonable that imputing missing values would have the most impact on the distances within that cluster. Interestingly, we see that cluster 5 did not change significantly despite adding observations from cluster 4 and imputing missing values.

D. CLUSTER ENSEMBLES

In this section, we show the results of generating an ensemble of clusterings. As outlined in Chapter IV, we use K-NN quantile sampling to randomly impute missing values 5,000 times. With the resulting 5,000 sets of data, we produce the Gower distances and PAM clustering for each sample. Then, we apply the clustering ensemble algorithm to compute the ensemble distances, and use these to map the 5,000 clusterings to two-dimensional space using classical dimensional scaling, as displayed in Figure 10.

Figure 10. Ensemble Clusterings Mapped to Two Dimensional Space for Visualization

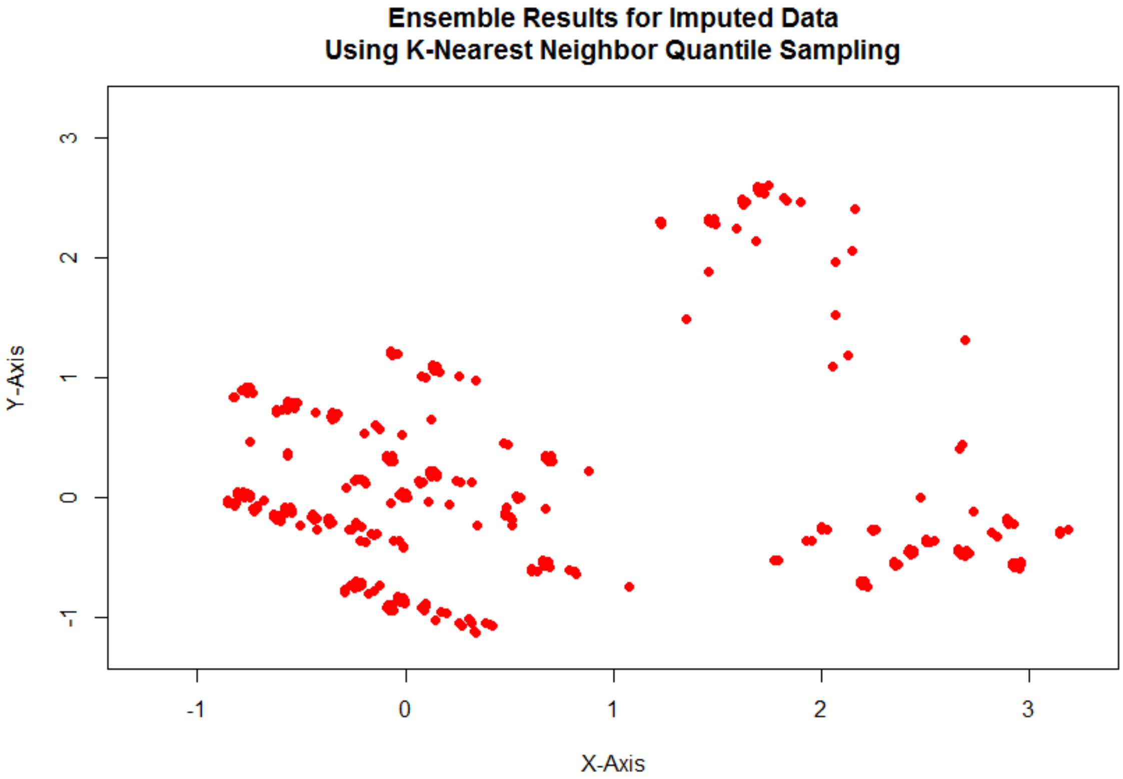


Figure 10 illustrates the ensemble of clusters mapped to a two-dimensional space, where each point represents a clustering.

In Figure 10, each point represents a clustering. And, using K-NN quantile sampling, we see that two primary groups of clusterings form, separating the data approximately where the x-axis is equal to 1.10. This is an important distinction because we also find that only 642 of the 5,000 observations have a horizontal axis value greater than 1.10, which suggests that the majority of the clusterings resides are "closer" together. However, this display of the data suggests that certain clusterings can be in disparate locations and produce differing results. For example, at the maximum point along the x-axis (x=3.18), we notice some non-trivial differences from how the cities cluster in Table 7. Ho Chi Minh City and Toronto both cluster with the Indian cities (cluster 3), Buenos Aires clusters with the western cities (cluster 1), and São Paulo forms a cluster with only Tehran, Bangkok, Jakarta, and Istanbul. This result illustrates how significant changes in the data can influence the clusterings in extreme cases.

With the understanding of what can occur in extreme instances, we then use the data resulting from the ensemble to cluster the cities into a consensus or average clustering. And, we outline results for both hard and soft consensus clusterings. The hard consensus clusters can be observed in Table 9.

Table 9. Consensus Clustering using a Cluster Ensemble of K-NN Quantile Sampling of Imputed Data and c=5.

Cluster 1 (10)	Cluster 2 (6)	Cluster 3 (6)	Cluster 4 (9)	Cluster 5 (10)
New York Los Angeles Chicago Washington, DC Dallas-FW Philadelphia San Francisco Boston Toronto London	Beijing Tianjin Shanghai Chongqing São Paulo Rio de Janeiro	Delhi Mumbai Kolkata Bangalore Hyderabad Johannesburg	Lagos Jakarta Bangkok Manila Moscow Istanbul Buenos Aires Lima Ho Chi Minh City	Cairo Kinshasa Karachi Al-Riyadh Tehran Kabul Dhaka Nairobi[4] Baghdad[4] Mexico City[4]

Table 9 shows the changes in cluster from the baseline data when we incorporate an ensemble of clusters. We see that the consensus cluster from our ensemble more closely represents our baseline cluster than one randomly generation repetition imputation of missing data.

Because the resulting hard clusters are very similar to those in our baseline clusters (92.7% consistent), we do not analyze the results in detail. But, they indicate that the clusters are generally stable in the long run with the exception of Nairobi, Baghdad, and Mexico City. Mexico City represents the most interesting shift among these cities because it possessed the largest silhouette width among the four that shifted as well as a larger value than Lagos or Moscow in the baseline clusters. For Moscow, this result also shows the difference between imputing the median of the five nearest neighbors versus the long run consensus cluster. When we use the median of its five nearest neighbors, we see that Moscow clusters with the western cities in cluster 1. But, when we allow for quantile sampling, it clusters more closely with the cities in cluster 4.

Given the resulting graph in Figure 10 and the subsequent changes due to data imputation, it is unlikely that we can truly force each of these cities into solely representing one cluster. Cities like Mexico City and Moscow that are known to switch clearly possess properties that make them good candidates to shift to their neighboring clusters. Therefore, we find it useful to illustrate this by examining the soft consensus clustering method for ensembles because it allows for weighted clusters. We expect observations with weights close to 1 to be strongly associated with the highly weighted cluster. And conversely, we expect observations with close or equal to zero weight in a cluster to have virtually no association with that cluster. Weights between 0.10 and 0.90 that spread across multiple clusters suggest that the observation is sensitive to changes in the data. For example, a city with equal weight of 0.50 spread across two different clusters indicates a level of indifference between the two clusters. In Table 10, we show the cities that possess at least some level of uncertainty in their clustering. The remaining cities all have a cluster weight of 1.0 and are associated with their cluster from the consensus clustering in Table 9. And, we denote the dominant cluster for each of the cities in this group by highlighting the weight applied to its dominant cluster in yellow.

Table 10. Table of Soft Clusters Showing Weights to Each Cluster

City	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Toronto	0.89	0.00	0.11	0.00	0.00
Lagos	0.00	0.00	0.00	0.74	0.26
Cairo	0.00	0.00	0.00	0.01	0.99
Nairobi	0.00	0.00	0.00	0.06	0.94
Karachi	0.00	0.00	0.01	0.05	0.94
Baghdad	0.00	0.00	0.00	0.02	0.98
Tehran	0.00	0.00	0.01	0.09	0.90
Jakarta	0.00	0.00	0.00	0.96	0.04
Manila	0.00	0.00	0.00	0.90	0.10
Mexico City	0.00	0.00	0.00	0.15	0.85
São Paulo	0.22	0.77	0.00	0.01	0.00
Moscow	0.32	0.00	0.00	0.68	0.00
London	0.97	0.00	0.00	0.03	0.00
Istanbul	0.00	0.00	0.22	0.78	0.00
Buenos Aires	0.10	0.00	0.00	0.90	0.00
Lima	0.00	0.00	0.00	0.87	0.13
Ho Chi Minh City	0.00	0.00	0.05	0.94	0.01
Rio de Janeiro	0.11	0.89	0.00	0.00	0.00

Table 10 only displays the cities for this study that do not have a weight of 1.0 in a single cluster. Consider this a table of uncertainty for these cities. In essence, the weights provide an indication of how strongly each observation associates with the cluster listed in the column. You can see Moscow has the weakest association with its dominate cluster.

The majority of observations in Table 10 are closely associated with one cluster and significantly less associated with others. But, it is important we note that some level of uncertainty exists. We focus on the cities with the most uncertainty in their cluster. São Paulo is more strongly associated with cluster 2 and has a secondary association with the western cities in cluster 1. Lagos, Istanbul, and Moscow more closely associate with cluster 4 but clearly have strong secondary associations to clusters 5, 1, and 3, respectively. Furthermore, if we compare the results of the hard clustering in Table 9 with the soft clustering in Table 10, we see that none of the observations associated with cluster 4 have a weight of 1. This confirms the intuition behind our results comparing the K-NN median sample to the base sample. Approximately 34% of the missing values form our data set come from cluster 4. Hence, we confirm the cities in that cluster are more susceptible to deviations than others due to their missing data. Moreover, cluster 4 is the only cluster where no observations have a weight of 1.0.

E. ENSEMBLE CLUSTERING COMPARISON

To better illustrate the differences between the 5,000 K-NN quantile sample clusterings, our consensus clustering, the K-NN median clustering, and our base clustering, we generate the ensemble distances between each of them and use classical multi-dimensional scaling to show them on a 2-D plot. We note that, for exploratory purposes, we also include an example where the imputed values are sampled from triangular distributions approximating the marginal distribution of the corresponding variable. First, we generate an ensemble of the 5,004 different clusterings. Then, we use classical multidimensional scaling to map them into the plot shown in Figure 11.

Figure 11. Comparison of 5,000 K-NN Quantile Samples to Base Clustering, Example Imputed Data Clusterings, and Consensus Clustering.

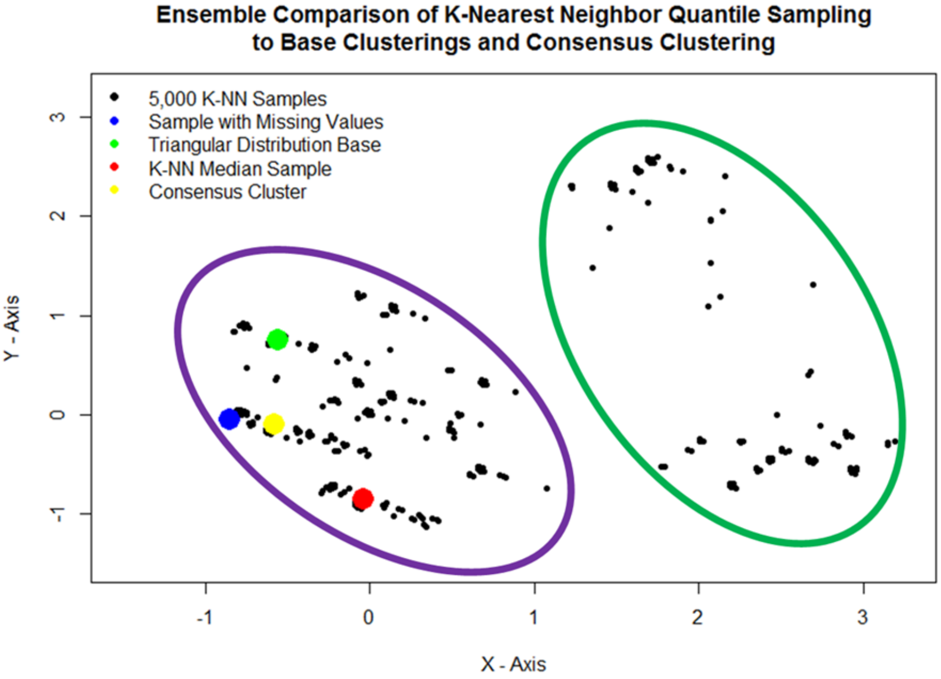


Figure 11 compares the results from earlier in Chapter V to our ensemble of clusterings and our consensus clustering. As shown, we see that the consensus cluster (blue) and base cluster (yellow-with missing values) are very similar. We also see that all base clusterings and the consensus cluster reside in the group of sample clusterings to the left outlined in purple, which consists of over 4,300 of the 5,000 samples.

The results in Figure 11 show us that the clustering results from using Gower's distance with using missing values are not very different from the consensus clustering. And in fact, Gower's distance with missing values is closer to the consensus clustering than using the K-NN median result or the triangular distribution example. We also notice that each of the three clusterings produced using other methods and the consensus cluster fall within the same generally area in Figure 11 as the group of K-NN quantile samples outlined in purple. This suggests that the clusterings outlined in green are more unlikely events and do not reflect what we can most likely expect to occur in our clustering. Moreover, it demonstrates how well Gower's method performs even when using a data set with 8–10% of its values missing. Clearly, translating observations into clusters using Gower's method to calculate the distances is relatively robust to missing data.

F. CONSENSUS CLUSTERING MEANS

The consensus clustering allows us to see how the cities cluster together on average, but it does not necessarily provide details of how the data is structured to produce those values. In order to do this, we use the results of the K-NN quantile sampling and the classical multidimensional scaling to see which data set is most similar to the consensus cluster. In our simulation of 5,000 runs of K-NN quantile sampling, 968 data sets produce a clustering result that covered the same two dimensional space as the consensus clustering (x=-0.580, y=0.105). To illustrate an example, we extract an example data set from our list and produce parallel coordinate plots of the average values for each variable by cluster. In this case, we use the average value in order illustrate what we can expect from the types of cities that associate with each cluster. And, because it shows only five lines, we can more clearly delineate the relationships between each cluster vice each individual observation. We generate the parallel coordinate plots (Venables & Ripley, 2002) in the same combinations of variables we use for the missing data results. Figure 12 shows the transportation network variables, Figure 13 shows access to utilities, Figure 14 shows government services, Figure 15 shows economic data, and Figure 16 shows demographic data.

Figure 12. Consensus Clustering Mean Transportation Network Values Comparison.

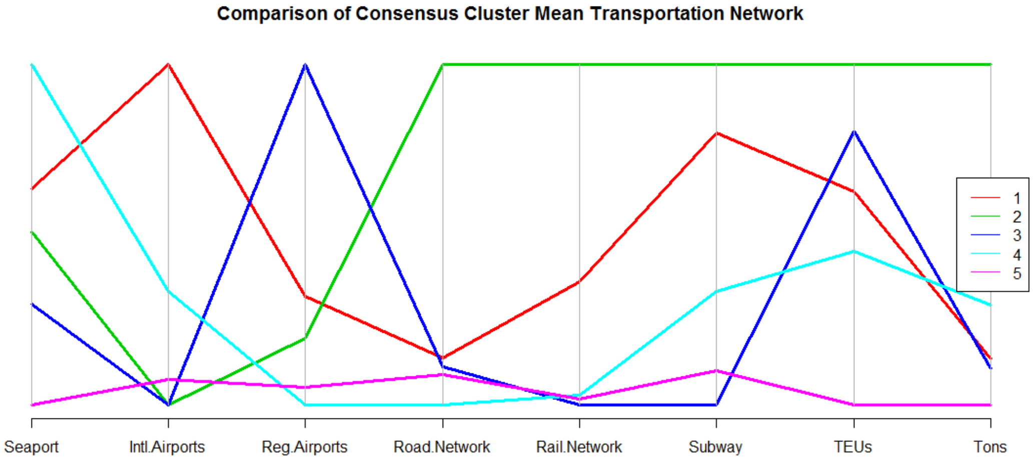


Figure 12 illustrates that most cities in cluster 4 have a seaport and robust international airport networks, but they lack robust road and rail networks. The cities from cluster 2 tend to have large ground and water transportation networks but less robust air transportation networks. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

Figure 12 shows what we can expect to observe from the cities in each cluster, on average, for their transportation networks when we account for imputed data. The transportation network data did not contain many missing values, so we see that the results closely resemble those from our baseline data. We see that cities such as Nairobi, Kinshasa, and Bagdad in cluster 5 tend to have less robust transportation networks across all transportation variables. Similarly, cities in cluster 4 like Lagos, Jakarta, and Istanbul tend to have less robust transportation networks as well. But, they tend to differentiate themselves by having access to a seaport. In fact, cluster 4 clearly had the most cities with seaport access. In contrast, the cities in cluster 2, the Chinese and Brazilian cities, tend to have more robust ground transportation networks but less access to air transportation. Again, we see that the cities in cluster 3, mostly from India, tend to have large networks of regional airports and throughput through seaports in TEUs despite a limited number of cities with seaport access. The western cities in cluster 1 clearly dominate international air transportation and have robust public transportation networks in the form of subways.

Figure 13. Consensus Clustering Mean Utility Network Values Comparison.

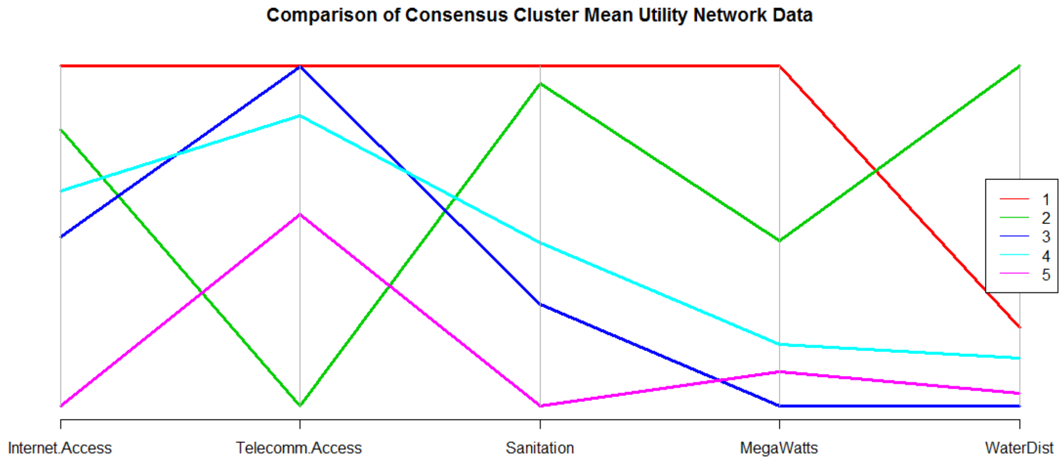


Figure 13 compares access to various utility networks among the five clusters. On average, the western cities clearly have the most access to utility services which makes sense. Interestingly, the cities in cluster 2 (Chinese and Brazilian), have good access to utilities as well with the exception of telecommunications. Clusters 3, 4, and 5 show a common aspect of the developing world – good access to telecommunications but poor access to sanitation, water, and power. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

We observe some common global trends regarding utility access in Figure 13. Western cities tend to be highly connected in this regard, which is not a particularly interesting result. However, when we look at the average trends for clusters 3, 4, and 5, we see a common component of the developing world. The cities in these clusters tend to have improved access to telecommunications over Internet access. However, they also tend to be characterized by poor access to basic sanitation, water needs and electricity. This is valuable in building our intuition about these large urban environments. For cities like these, we can expect them to lack robust infrastructure to provide these basic services to their citizens. Specific to cluster 4, when we couple this with the large percentage of Pacific Rim nations, we see that challenges with natural disasters such as typhoons or tsunamis can pose significant problems to the provision of basic utilities. In the context of combat operations, we also see that cities in clusters 3, 4, and 5 will also have broad access to the Internet and (mobile) telephones for messaging and information operations.

Figure 14. Consensus Clustering Mean Government Service Values Comparison.

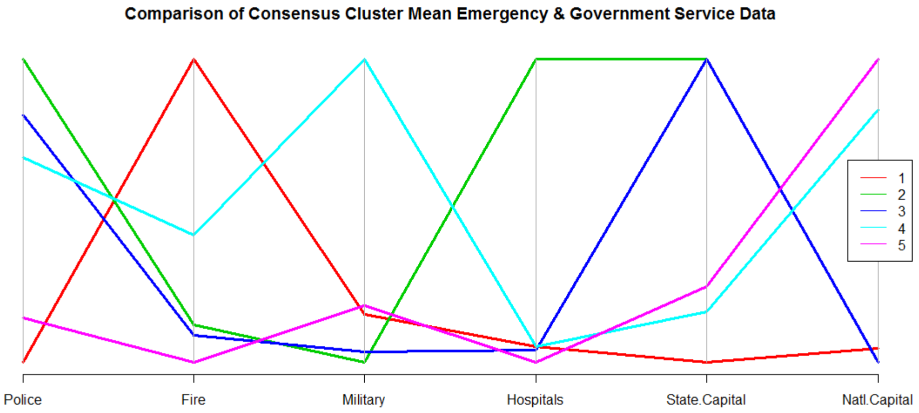


Figure 14 displays the average availability of government services for each cluster in the consensus cluster. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

The results in Figure 14 do not change significantly from the results in the base data section, in part because government services did not have large number of missing values.

Figure 15. Consensus Cluster Mean Economic Data Values Comparison.

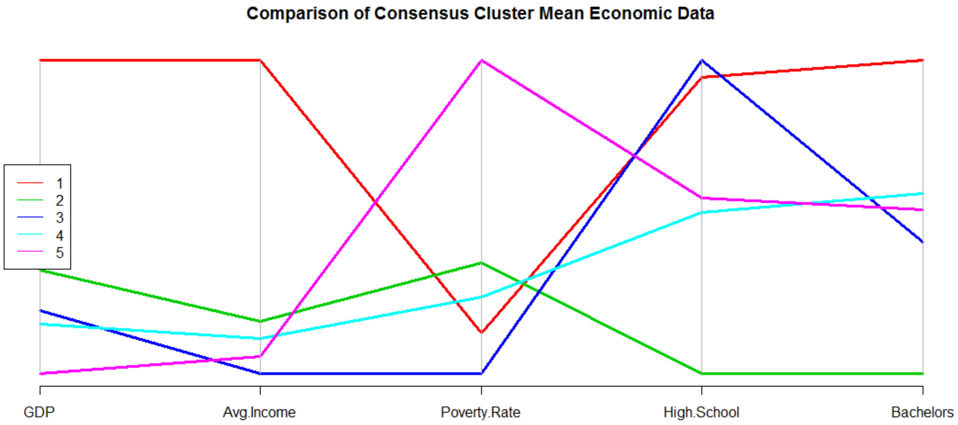


Figure 15 shows the average results for economic data in the consensus cluster. Cluster 5 has high poverty rates. Clusters 2–5, all have low GDPs and average household incomes relative to the western cities in cluster 1, on average. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

Among the economic variables, education (High.School and Bachelors) contains the most missing values in the base data. In Figure 8 from Chapter IV, we see that clusters 4 and 5 contain a large number of missing values. And, the values that are available tend to be low. The results in Figure 15 suggest that, on average, the values for education tend to be higher with imputed data. This may indicate that more research is required into the education data for those cities with missing values. The variable for poverty rates also contained a large number of missing values, but we notice that the results do not change significantly from the base results. Cities in cluster 5 such as Kinshasa, Tehran, and Nairobi, tend to have high poverty rates among their populace and low purchasing power in the form of average annual income.

Figure 16. Consensus Clustering Mean Demographic Data Values Comparison.

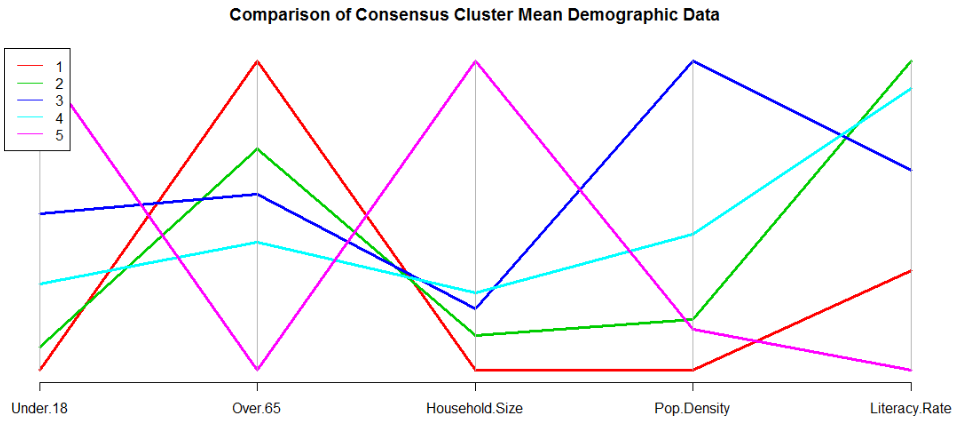


Figure 16 compares demographic data for each of the clusters. This depiction clearly displays that cities in the developed world have older populations, smaller household sizes, and lower population densities, whereas cities in the developing world tend to have young populations, large household sizes and dense city populations. Sources used for compilation can be found in Appendix C, and plotted using techniques from Venables & Ripley (2002).

The depiction in Figure 16 clearly articulates the importance of accounting for dependence when imputing missing values. We can clearly see that cities with a high percentage of young people tend to possess a low percentage of older people and vice versa. Interestingly, however, the cities from clusters 3 and 4 tend to be relatively balanced. As we might expect, we also see that the developing world cities in cluster 4 and 5 tend to have large household sizes. These are important distinctions to make between cities in terms of human geography. Generally, the remaining variables are consistent with our base results.

G. COMPARISONS TO PREVIOUS WORK

Generally, our work compares favorably to the megacity work done previously in United States Army (2014) and Sapol (2016). First we compare the results of our work to that of United States Army (2014). Our results generally produce similar results, but we are able to do so with many more cities as well as more variables. This allows us to produce a more rich and diverse set of results. Table 11 compares our consensus cluster to those of United States Army (2014) where HC stands for highly connected, MC stands for moderately connected, and LC stands for loosely connected as depicted in United States Army (2014).

Table 11. Comparison of K-NN Consensus Clustering with U.S. Army Case Studies from United States Army (2014).

Cluster 1 (10)	Cluster 2 (6)	Cluster 3 (6)	Cluster 4 (9)	Cluster 5 (10)
New York (HC)	Beijing	Delhi	Lagos (LC)	Cairo
Los Angeles	Tianjin	Mumbai	Jakarta	Kinshasa
Chicago	Shanghai	Kolkata	Bangkok (MC)	Al-Riyadh
Washington, DC	Chongqing	Bangalore	Manila	Kabul
Dallas-FW	São Paulo (MC)	Hyderabad	Moscow	Karachi
San Francisco	Rio de Janeiro (MC)	Johannesburg	Istanbul	Tehran
Philadelphia			Buenos Aires	Dhaka (LC)
Boston			Lima	Nairobi
Toronto			Ho Chi Minh City	Baghdad
London				Mexico City

Table 11 shows our consensus cluster using the K-NN quantile sampling, along with the groupings from United States Army (2014) using a two letter code where HC signifies the city is highly connected, MC signifies the city is moderately connected, and LC signifies the city is loosely connect in United States Army (2014). Note the cities from our study that cluster with each city from United States Army (2014), given clustering groups cities that are most similar given a set of variables and number of clusters.

As seen in Table 11, the cities with different levels of connectedness from United States Army (2014) generally do not cluster together in our study either. However, we see that Rio de Janeiro and São Paulo cluster together in both studies. Interestingly, Lago and

Dhaka were both loosely connected (LC) in United States Army (2014), but they are in clusters 4 and 5 in our study. And, Bangkok is moderately connected in United States Army (2014) but clusters with cities like Lagos in our study. This represents an interesting and useful result. When we can add more cities and more variables for comparison, we see that some cities that appear similar may not be, and conversely, some cities may actually be similar that do not appear so on the surface. In United States Army (2014), it appears as though Bangkok is more similar to São Paulo and Rio de Janeiro than to Lagos. However, we show that accounting for more detailed quantitative data on infrastructure, government services, economic development, and demographics draws Lagos and Bangkok closer together. Furthermore, when we refer back to Table 10 showing the soft clusterings, we see that Bangkok has virtually no weight (0.01) associated with cluster 2 where São Paulo and Rio de Janeiro reside. And conversely, Rio de Janeiro's weight in cluster 4 is 0.01, signaling it has virtually no association with the cities in that cluster on average. We expect the results that produced those very small values were extreme points and not representative of likely outcomes.

Next, we compare the results of our study to those found in Sapol (2016). Recall, Sapol (2016) uses only two variables, GDP per capita and population density, to cluster megacities into six different groups. Table 12 shows the comparison between our results. We again use the cluster memberships from our consensus clustering, and we annotate the cities from Sapol (2016) with the numerical value of their tier in parentheses.

Table 12. Comparison of K-NN Consensus Clustering with Megacity Classification Framework from Sapol (2016).

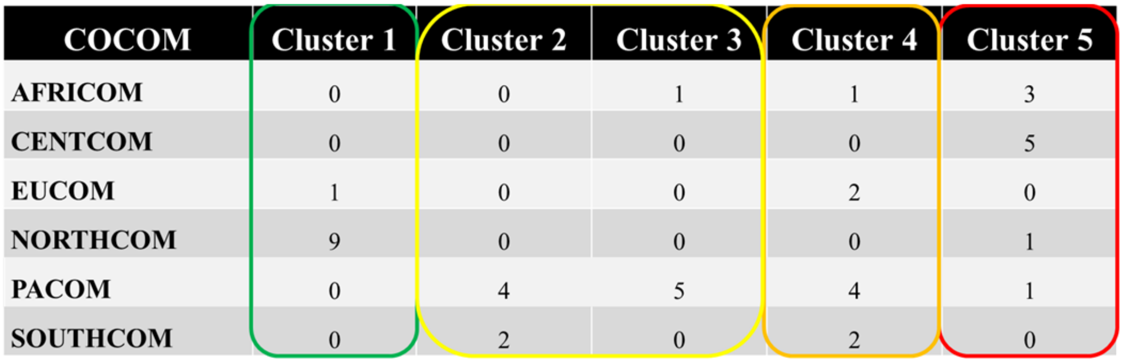
Cluster 1 (10)	Cluster 2 (6)	Cluster 3 (6)	Cluster 4 (9)	Cluster 5 (10)
New York (1) Los Angeles (1) Chicago Washington, DC Dallas-FW San Francisco Philadelphia Boston Toronto London (1)	Beijing (3) Tianjin (3) Shanghai (3) Chongqing São Paulo (3) Rio de Janeiro (3)	Delhi (4) Mumbai (6) Kolkata (5) Bangalore (5) Hyderabad Johannesburg	Lagos (5) Jakarta (5) Bangkok (3) Manila (4) Moscow (2) Istanbul (4) Buenos Aires (3) Lima (4) Ho Chi Minh City (5)	Cairo (5) Kinshasa (6) Al-Riyadh Kabul Karachi (6) Tehran (5) Dhaka (6) Nairobi Baghdad Mexico City (4)

In Table 12, we identify the corresponding cluster from Sapol 2016. We note that he uses six clusters instead of five, but the majority of cities from his cluster 2 are not in this study with the exception of Moscow. We see that generally clusters 5 and 6 from his study coincide with clusters 4 and 5 from our study aside from Mumbai, Kolkata, and Bangalore. And, our cluster 2 corresponds well with cluster 3 in Sapol (2016).

While there are similarities between our consensus clustering and the clustering from Sapol (2016), we notice that additional variables and imputed data cause cities to shift and associate differently. One interesting result is the shift in Moscow’s cluster. In Sapol (2016), Moscow clusters with cities like Tokyo, Seoul, and Nagoya. We do not include those cities in our study, so we may expect that Moscow would shift closer to them if they are added. However, this is uncertain given they would be added in the presence of the additional variables we incorporate into our study. Future work can look at adding these cities as well as cities like Lahore, Pakistan or Paris, France.

Sapol (2016) also makes the case that we should examine transitioning from a regional focus to a megacity focus given the importance of megacities in the future operating environment (FOE). We consider this in our analysis and generate a table of our results based on the major combatant commands (COCOM) which can be seen in Table 13. We do not show the names of each city but rather include the number of cities associated with each COCOM by cluster. We also apply red, orange, yellow and green outlines to align our clusters with the tiers from Sapol (2016) and the levels of connectedness from United States Army (2014), where green is highly connected (HC), yellow is moderately connected (MC), and orange and red are generally loosely connected (LC).

Table 13. Count of Cities Aligned with Each U.S. COCOM by Cluster.



COCOM	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
AFRICOM	0	0	1	1	3
CENTCOM	0	0	0	0	5
EUCOM	1	0	0	2	0
NORTHCOM	9	0	0	0	1
PACOM	0	4	5	4	1
SOUTHCOM	0	2	0	2	0

Table 13 shows the number of cities in each cluster for each COCOM. We compare them to tiers of Sapol (2016) and United States Army (2014) using the green, yellow, orange, and red outlines such that green is HC, yellow is MC, and orange and red are LC.

As shown in Table 13, we see that our clustering is indicative of the challenges facing commanders in the arc of instability. AFRICOM, CENTCOM, and PACOM dominate the cities that would be referred to as loosely connected in United States Army (2014) or tier 5 and tier 6 in Sapol (2016). PACOM clearly has the largest and most diverse problem set, including five cities considered to be loosely connected. These are likely to have higher incidences of organized crime, terrorism, and other crises that may require responses like humanitarian assistance and disaster relief due to their poor infrastructure and access to services. Additionally, they also face the dynamic megacity problem of 9 cities that United States Army (2014) may refer to as moderately connected. They come from more stable regions but still have infrastructure, economic, and social challenges that can lead to destabilization. In contrast, CENTCOM only has five cities in this study but still has a dynamic and challenging problem because they all correspond to the loosely connected tier and have high risks of instability. Notably, in 2017, U.S. forces remain in Afghanistan conducting operations in the country where Kabul is located; Baghdad is located in Iraq where operations continue to occur against ISIS; and Iran (Tehran) continues to have an adversarial relationship with the U.S. and present substantial foreign policy challenges. In the other COCOM with a dynamic problem set, AFRICOM, we see that four of the five cities align with our loosely connected or high risk of instability clusters. With factions of Al-Qaeda in Egypt and Boko Haram operating in Nigeria, we see that AFRICOM faces its own challenges in supporting and ensuring stability in the region.

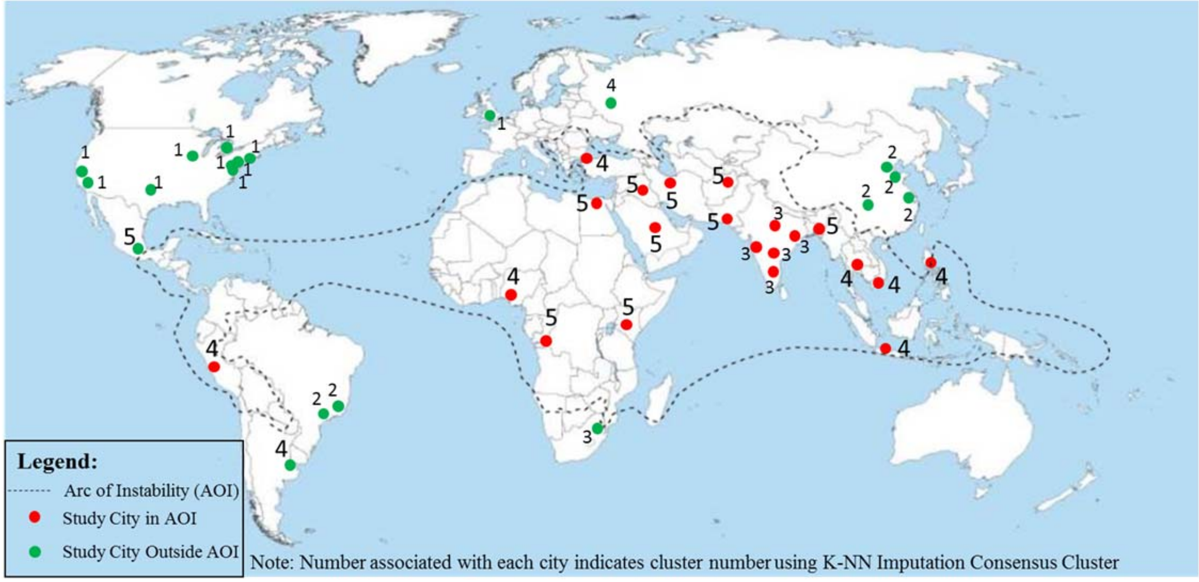
VI. CONCLUSIONS AND FUTURE WORK

In this study, we use 41 different large urban areas across the globe and 33 variables constructed from over 90 public sources to create clusters and identify what cities are similar or different in meaningful ways. With uncertainty in our data through 103 missing values, we use a combination of data imputation and cluster ensembles to identify the sensitivities of our clusters to data uncertainty. We establish baseline clusters using Gower distances and the PAM algorithm based on observations with missing values, and then we compare these results to clustering with missing values imputed using the K-NN median method. We find that imputing missing values does not significantly change the clusters and primarily impacts the cities with larger amounts of missing data and looser connections to their baseline clusters. Finally, we randomly impute values and apply techniques from Hornik (2005) using the CLUE package to create an ensemble of 5,000 clusterings. This ensemble of clusterings identifies the sensitivities of clusterings to missing data and allows us to produce both hard and soft consensus clusterings. Our findings indicate that the consensus or average clustering for our data does not differ greatly from our baseline clusters, which demonstrates the stability of our data set. These methods and our approach to constructing the 33 variables in our data set not only help inform JWAC of the similarities and differences between the 41 large urbans areas, they also provide an analytical approach for identifying for which cities more data is warranted. Furthermore, it provides an analytical framework for future work in this area.

A. CONCLUSIONS

As discussed in Chapters I and II, the arc of instability continues to shape our future security environment. It represents parts of the developing world where competition for resources, terrorism, political instability, and economic inequality make conditions ripe for conflict, which have pushed it to the forefront of American foreign policy decisions since the early 2000s (United States Marine Corps, 2015; Barnett 2004). Our results help inform some of the challenges of the arc of instability. The majority of cities with the poorest infrastructure, lowest access to utilities, and poorest economies group together in clusters 4 and 5. And, we find that many of the cities from our data set that are in the arc of instability cluster with other cities within the arc of instability. Or, if they are not located in the arc, we find that they tend to cluster with other cities on the outside. In order to see this clearly, examine Figure 17. We display a map with the 41cities from our study, along with the cluster to which they belong using our consensus or average clustering from the K-NN quantile sampling imputations.

Figure 17. Arc of Instability Map with Overlay of Cities and Consensus Clusters. Adapted from United States Marine Corps (2014).



Note: Number associated with each city indicates cluster number using K-NN Imputation Consensus Cluster

Figure 17 shows a point on the map for each city in our study as well as the cluster to which it belongs. Cities within the arc of instability are annotated in red, and cities outside the arc are annotated in green. As discussed in Chapter V, cities from clusters 4 and 5 generally reside within the arc of instability, along with cluster 3. We distinguish cluster 3 because India is not included in the AOI in some literature e.g. Barnett (2004) but is included in others e.g. United States Marine Corps (2015).

As we outline in Chapter II, Kilcullen (2013), United States Army (2014), and United States Marine Corps (2015), each make the case in different ways that the arc of instability will continue to be an important component to our future operating environment. And, cities that are located within the arc of instability face greater risks of requiring military action, either combat or HA/DR. However, we must also be cognizant of cities that are not in the arc of instability but cluster with cities in the arc. Specifically, we note that Johannesburg, Mexico City, Lima, and Moscow all reside outside the arc but are most similar to cities within the arc. Hence, there may be something that the data is revealing to us about the nature of these cities and their future risks. United States Army (2014) and Kilcullen (2013) make note of the reality that competition for resources, growing income inequality, urbanization, and other socio-economic and demographic factors can increase the likelihood of instability in a particular city or region. Therefore, cities such as those may warrant closer attention from intelligence professionals, planners, and senior leaders.

Our work also illustrates the benefits of combatant commands conducting information sharing and working in concert with one another on promoting stability. As displayed in Table 13, each cluster contains cities from multiple combatant commands. We know that these cities are, by definition, similar across their four pillars which suggests there may be events or dynamics occurring in one that may also drive future events in another. But, we do not advocate for a shift from the COCOM structure in favor of a megacity framework. Of the 41 cities we study, 21 are within the arc of instability and 10 of those reside solely in the PACOM area of responsibility. And, with the exception of PACOM, we notice cities in the other COCOMs do not face the same scope and complexity of their problem set. But, there may be value in potentially splitting PACOM into PACOM East/West or PACOM North/South. The area covered, swiftly growing populations, high potential for natural disasters, and competition for resources require a more detailed focus of effort that could warrant two separate COCOM structures. And, while we cannot predict with certainty the actual areas where instability or conflict will arise given current data limitations, we can further the discussion and identify the areas where effort can increase.

In comparison to the previous work done in United States Army (2014) and Sapol (2016), our work adds more quantitative data in terms of the number of cities and especially the number of variable. In the case of United States Army (2014), we are able to achieve similar results without the use of expensive temporary duty trips while simultaneously incorporating more large urban areas. Additionally, our work lays out a methodology for adding more cities to the work in the future. In contrast, United States Army (2014) may require site surveys, logistical support, or potentially special operations forces to gain deep insights into some of these large cities. This allows us to gain insights into cities without significant additional costs, which is particularly true for cities in countries where the U.S. has limited access. Our results also differ from the findings of Sapol (2016). We find that when we incorporate other factors beyond GDP per capita and population density, some of the clusters shift. While we note that some of the cities differ, we see that cities in India move from being in separate clusters in Sapol (2016) to being in the same cluster in our study. And, like Sapol (2016) cities in the arc of instability generally cluster together.

B. FUTURE WORK

Our study does not provide revolutionary insights into the megacity problem or capture detailed information regarding the interdependencies of megacity critical networks. These challenges are complex and require a substantial amount of data. We consider this a building block on previous studies and an effort to drive the conversation forward regarding the potential for future military operations in these large urban areas. As a result, we believe several different lines of effort exist for expanding this study. First, future researchers can leverage the data collection sources and techniques we use in this study to expand the number of variables and more accurately capture information. In previous chapters, we briefly highlight that data for access to emergency services does not adequately distinguish the cities into clusters. Future researchers can incorporate additional details like per capita number of beds, violent crime rates, or stations per 1000 people in order to more accurately capture true access and support.

Additionally, trade uses other means of transportation beyond waterways. Hence, researchers can look to adding throughput of passengers and freight through critical air and ground transportation nodes. While difficult to find in many cases, we also expect the daily or year throughput in dollars to have significance. Flights out or major airports being stopped, isolating main supply routes, and conducting naval blockades can all prevent the flow of goods and services into and out of a city. If we do this, the value of throughput may signal losses to the city’s economy due to our operations. At the same time, we expect primary industry is also important to understanding the general make up of a megacity. Cities whose primary industry is manufacturing will inevitably contain more plants and warehouses, which is important to understand when preparing for combat operations.

Military requirements and objectives also require attention from research of this nature. In our study, we select a combination of general data that pertains to our four pillars. However, combat operations or HA/DR support may require information pertaining to different types of infrastructure like radar sites or supply depots. Future work might solicit subject matter experts for their insights into the types of infrastructure and human geography data that will most benefit operations. While military specific data will likely prove difficult to collect, it will allow researchers to use the data and similar methodologies to ours in order to identify how cities cluster based on military requirements data.

Additionally, researchers can look to adding cities to the study or changing the data imputation technique. Additional cities will continue to fill the dimensional space and potentially allow us to use more clusters. With more clusters, we have the opportunity to allow the representative observations or medoids to have shorter distances between them and the other observations within the cluster. Moreover, additional cities grant us more visibility on regions of interest to senior leaders and JWAC. We also suggest examining different data imputation techniques that are more specific to each individual variable. Finally, we suggest that a simple interactive tool be developed that allows analysts to explore megacity data of the type we have constructed using techniques outlined in this thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. SUMMARY DATA

Table 14. Summary Data for Continuous and Binary Variables in the Data Set

Variable	Mean	Std. Deviation	Percent Missing
Electricity Generation (MW)	6073.58	5985.90	0.00
Ground Water Sources	1.88	5.75	17.50
Reservoir/River/Lake Sources	7.03	15.31	12.50
Desalination Plants	0.06	0.24	12.50
Daily Water Dist. (million gal.)	664.45	494.23	0.00
Int'l Airports	1.80	1.20	0.00
Regional Airports	1.10	1.65	0.00
Seaport	0.48	0.51	0.00
Seaport (million TEUs)	5.67	11.69	0.00
Seaport (million tons)	26.58	80.92	0.00
Road Network (miles)	6343.27	14640.35	12.50
Rail Network (miles)	193.70	260.45	12.50
Telecomm Access (%)	80.75	27.84	20.00
Subway	0.42	0.50	0.00
Sanitation Access (%)	73.90	29.02	7.50
Hospitals	163.26	264.80	2.50
Police Stations	60.62	41.82	0.00
Fire Stations	96.82	130.34	0.00
Military Bases	13.18	19.17	5.00
State Capital	0.50	0.51	0.00
National Capital	0.50	0.51	0.00
GDP (billion USD)	298.47	331.46	2.50
High School Diploma (%)	68.84	23.91	27.50
Bachelor's Degree (%)	24.93	14.39	32.50
Literacy Rate (%)	83.21	16.23	20.00
Average Income	25696.35	29113.56	12.50
Poverty Rate (%)	18.49	15.04	17.50
Average Household Size	3.63	1.25	2.50
Pop Under 18-19 Yrs Old (%)	28.51	9.70	7.50
Pop Over 65 Yrs Old (%)	8.93	4.71	10.00
Internet Access (%)	58.23	28.79	20.00
Pop Density (Per sq. mile)	12834.23	15777.65	0.00

Religion is not included in this table because it is a categorical variable. However, we note the counts for religion as follows: Christian (18), Islam (10), Hinduism (5), Chinese or Other Religion (4), Buddhism (1), and Vietnamese or Other Religion (1).

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. VARIABLE ESTIMATION

In an optimal study scenario, we have access to data that perfectly estimate the desired variables. However, in practice, we recognize that this is often infeasible and variables will often contain intrinsic uncertainty. In order to establish baseline estimates for our study, many of the variables required proxies and additional calculations in order to obtain a reasonable level of information about the variables to which we sought. In the following sections of Appendix B., we outline the techniques and proxies we used to produce our variable estimates for those variables that required more approximation. And in general, we clarify that we make approximations for unavailable data cities using state or provincial level data when the other methods from this appendix do not produce results.

A. ELECTRICITY GENERATION

For electricity generation, we required a combination of proxies to get estimates and recognize these methods contain some inherent weaknesses. We sought data primarily on the amount of installed capacity (MW) in each city, but this information was not available for some of the cities in our study. As a result, we used electricity generation or electricity consumption as proxy variables for installed capacity in cases where the desired data was not available. We also encountered cities that had per capita level consumption data from previous years but no recent data. In these cases, we used the per capita figures and the current population to obtain an estimate for current levels of consumption.

B. WATER SOURCES AND DISTRIBUTION

We count one ground water aquifer listed by a city as a ground water source. And, we use a summation of the total number of lakes, other reservoirs, and rivers as the number of lakes/reservoirs/rivers variable. In the case of desalination, the majority of cities either do not currently have one or theirs will not become operational for the next few years. Hence, presence of a desalinations plant acts effectively as a binary variable. Sourcing data regarding water distribution required more indirect information. For example, if a city did not identify their daily water distribution, they may indicate the per capita consumption per day or the daily water demand and some percentage of access. In the case of per capita consumption, we use the cities current population estimate to calculate the daily water distribution. And, in the case of percent access and demand, we simply multiple demand by the percent access. For clarity, we also not that we use water supplied instead of water demand in cases there the two numbers disagree.

C. AIRPORTS

For international and regional airports, we consider any airport that conducts flights out of the home country as international, and any airports that only operate within the country as regional.

D. SEAPORTS

For clarity, we note that we do not consider river ports unless they open directly to the ocean as is the case with London's seaport. Additionally, we consider total throughput for both TEUs and metric tons.

E. ROAD, RAIL NETWORKS, AND SUBWAYS

For road networks, we include all paved roads and highways for the data that is available. In cases where the data is not available for both, we include only the data available without any estimation. In contrast, if the only data available is the road density, we convert this to the number of miles of road using the square area of the city. For rail networks, we have limited information regarding the amount of national or state rails that move through the city. Additionally, some of these cities include different types of rail networks including light rails, metros, subways, and heavy rails. For the purposes of this study we include data from all available rail networks.

F. TELECOMMUNICATIONS AND INTERNET ACCESS

We combine these basic they face similar data collection challenges. We calculate these as the percentage of the population with access to a phone (landline or mobile) and the percent with access to the Internet. For cities that do not provide this information directly, we use the number of subscriptions or users per 100 people. We use this as a proxy for percentage except in cases where the number of users or subscriptions per 100 people are above 100. In those cases, we default to an assumption of full penetration and use a value of 99.00% as the variable input.

G. GDP/GRDP AND AVERAGE INCOME

Some cities did not contain readily available and current GDP and average income data. However, some contain per capita GDP or income. For per capital GDP we multiply this by the current population and where growth rates are not available. For clarity, we only use figures listed in current prices.

H. POVERTY RATE

We recognize that different countries determine what constitute poverty relative to their location standards of living. As a result, this data provides some challenges in determining similarities or differences. Where available, we use the percentage of persons living on $1 or less per day because it is universal. This was most common in African and Asian countries. Where we cannot find suitable data in that regard, we shift to using the local poverty rate where available.

I. AGE DISTRIBUTION

In order to estimate the age distribution for each city in our study, we required proxy variables for young people in some cases. Generally, the U.S. cities were the only ones that used under 18 and over 65 as part of their age distribution. In contrast, most foreign cities do not use this metric. Their data sources use under 19 for estimating young people. Given the small differential, we use 19 and under as a good proxy for the percent of young people in cities.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. DATA SOURCES

Variable	Country	Source
Electricity Generation (MW)	Afghanistan	Power Plans Around the World, 2016
	Argentina	Buenos Aires City, 2010
	Bangladesh	Power Grid Company of Bangladesh, 2016
	Brazil	Ministry of Mines and Energy, 2015
	Canada	Aplin, 2013
	China	China Data Online, 2017
	Democratic Republic of the Congo	UN Data, 2017
	Egypt	Cairo Electricity Production Company, 2017
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	Tehran Times, 2016
	Iraq	Kneoma, 2017
	Kenya	Open Data for Africa, 2017d
	Mexico	Government of Mexico, 2017
	Nigeria	Open Data for Africa, 2017a
	Pakistan	Kneoma, 2017
	Peru	Ministry of Energy and Mines, 2015
	Philippines	Philippine Government, 2016
	Russia	Kneoma, 2017
	Saudi Arabia	Kingdom of Saudi Arabia, 2017
	South Africa	Eskom, 2017
	Thailand	Metropolitan Electricity Authority, 2014
	Turkey	Hurriyet Daily News, 2014
	United Kingdom	Department of Business, Energy & Industrial Strategy 2016
	United States	United States Energy Information Administration, 2017
	Vietnam	Do, 2002

Variable	Country	Sources
Ground Water Sources Reservoir/River/Lake Sources Desalination Plants	Afghanistan	Campbell, 2015
		USAID, 2016
		USGS, 2009
	Argentina	Buenos Aires City, 2010
		Engel, et al., 2011
	Bangladesh	
	Brazil	Tortajada, et al., 2006
	Brazil	National Water Agency, 2010a
	Brazil	National Water Agency, 2010b
	Canada	Toronto, 2017b
	Canada	Toronto, 2017c
	China	China Data Online, 2017
		People’s Republic of China, 2009
		University of British Columbia, 1999
		Probe International Beijing Group, 2008
	Democratic Republic of the Congo	Open Data for Africa, 2017e
	Egypt	Open Data for Africa, 2017b
	Egypt	Tour Egypt, 2017
	India	Indiastat, 2017
		Delhi Government, 2015
		BCPT, 2017
		Chaterjee, 2014
		BWSSB, 2016
		HMWSSB, 2015
	Indonesia	BPS, 2017
		Tutuko, 1998
	Iran	Nasseri, 2014
	Iraq	Kneoma, 2017
		Barbooti, et al., 2010
	Kenya	Karanja, 2011
		Nairobi City Water and Sewerage Company, 2017

Variable	Country	Sources
Ground Water Sources Reservoir/River/Lake Sources Desalination Plants	Mexico	Tortajada, et al., 2006
	Nigeria	Lagos State Government, 2013
	Pakistan	Karachi Water & Sewerage Board, 2013
		DHA Cogen LTD, 2017
	Peru	LIWA, 2014
	Philippines	Republic of the Philippines, 2017
	Russia	Mosvodokanal, 2017
		Landing and Planning, 2017
	Saudi Arabia	Kingdom of Saudi Arabia, 2017
		Tortajada, et al., 2006
	South Africa	Johannesburg Water, 2017
	Thailand	Kneoma, 2017
	Turkey	Turkey Statistical Institute, 2017b
		Demirci and Butt, 2001
	United Kingdom	Greater London Authority, 2017
	United States	New York City Government, 2015
		New York State, 2017
		The City of Newark, 2015
		City of Chicago, 2017b
		Los Angeles, 2017
		Orange County Water District, 2016
		MWRA, 2017 and BWSC, 2017
		City of Philadelphia, 2017b
		San Francisco Public Utilities Commission, 2016
		East Bay Municipal Utility District
		Texas Water Development Board, 2014
		City of Fort Forth, 2017b
		City of Dallas, 2017
		Texas Water Development Board, 2017
		DC Water, 2017
		Arlington County, 2015
		Virginia American Water, 2017
		WSSC, 2017
		Fairfax Water, 2017
	Vietnam	Institute for Global Environmental Studies, 2007

Variable	Country	Sources
Int’l Airports Regional Airports	Afghanistan	Google Maps Query
	Argentina	Google Maps Query
	Bangladesh	CAA, Bangladesh, 2006
	Brazil	Google Maps Query
	Canada	Google Maps Query
	China	Ministry of Commerce, 2007
		Google Maps Query
	Democratic Republic of the Congo	Google Maps Query
	Egypt	Google Maps Query
	India	Indiastat, 2017
	Indonesia	Google Maps Query
	Iran	Tehran, 2017
	Iraq	Google Maps Query
	Kenya	Nairobi Government, 2017d
	Mexico	Google Maps Query
	Nigeria	Google Maps Query
	Pakistan	Google Maps Query
	Peru	Google Maps Query
	Philippines	Google Maps Query
	Russia	Google Maps Query
	Saudi Arabia	Google Maps Query
	South Africa	Google Maps Query
	Thailand	Google Maps Query
	Turkey	Istanbul Government, 2017
	United Kingdom	Google Maps Query
	United States	United States Department of Transportation, 2017
	Vietnam	Google Maps Query
		Google Maps Query

Variable	Country	Sources
Seaport Seaport (million TEUs) Seaport (million tons)	Afghanistan	Google Maps Query
	Argentina	Buenos Aires City, 2010
		Buenos Aires Port, 2016
	Bangladesh	Google Maps Query
	Brazil	Google Maps Query
		Rio Port Authority, 2014
		Rio Port Authority, 2016
	Canada	N/A
	China	China Data Online, 2017 & IAPH, 2015
	Democratic Republic of the Congo	Google Maps Query
	Egypt	Google Maps Query
	India	Indiastat, 2017
		Mumbai Port Trust, 2017
	Indonesia	BPS, 2017 & IAPH, 2015
	Iran	Google Maps Query
	Iraq	Google Maps Query
	Kenya	Google Maps Query
	Mexico	Google Maps Query
	Nigeria	Open Data for Africa, 2017a
		Nigerian Ports Authority, 2015
	Pakistan	Karachi Port Trust, 2016
	Peru	Peru National Port Authority, 2016
	Philippines	Philippines Ports Authority, 2016
	Russia	N/A
	Saudi Arabia	Google Maps Query
	South Africa	Google Maps Query
	Thailand	ASEAN Ports, 2001
		IAPH, 2015
	Turkey	MARDAS, 2012 & World Port Source, 2017
	United Kingdom	Port of London Authority, 2014
		Port of London Authority, 2017
	United States	United States Department of Transportation, 2015
	Vietnam	Vietnam Seaports Association, 2015a
		Vietnam Seaports Association, 2015b

Variable	Country	Source
Telecommunications Access (%)	Afghanistan	Kneoma, 2017
	Argentina	Buenos Aires City, 2010
		Koop, 2015
	Bangladesh	BTRC, 2015
		Bangladesh Population Estimate 2015
	Brazil	N/A
	Canada	N/A
	China	China Data Online, 2017
	Democratic Republic of the Congo	UN Data, 2017
	Egypt	N/A
	India	Telecom Regulation Authority of India, 2017
	Indonesia	BPS, 2017
	Iran	N/A
	Iraq	Kneoma, 2017
	Kenya	Open Data for Africa, 2017d
	Mexico	INEGI, 2017
	Nigeria	N/A
	Pakistan	National Institute of Population Studies, 2013
	Peru	National Institute of Statistics and Information, 2015
	Philippines	N/A
	Russia	Kneoma, 2017
	Saudi Arabia	N/A
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	N/A
	United Kingdom	Kneoma, 2017
	United States	United States Census Bureau, 2015b
	Vietnam	Skuse, n.d

Variable	Country	Sources
Road Network (miles)	Afghanistan	N/A
	Argentina	Kneoma, 2017
	Bangladesh	Mahmud, 2014
	Brazil	Biderman, 2008
	Brazil	TomTom International, 2016
	Canada	Toronto, 2012
	China	Ministry of Commerce, 2007
		China Data Online, 2017
	Democratic Republic of the Congo	UN Data, 2017
	Egypt	Open Data for Africa, 2017b
	India	Indiastat, 2017
	Indonesia	Jakarta Government, 2014
	Iran	Kneoma, 2017
	Kenya	N/A
	Mexico	Mexico City, 2009
	Nigeria	Open Data for Africa, 2017a
	Pakistan	Hussain, 2011
	Peru	Moreira, et al., 2013
	Philippines	DPWH, 2016
	Russia	N/A
	Saudi Arabia	High Commission for the Development of ArRiyadh
	South Africa	City of Johannesburg, 2017
	Thailand	N/A
	Turkey	KGM, 2017
	United Kingdom	Kneoma, 2017
	United States	The City of Chicago, 2017a
		The City of Los Angeles, 2008
		The City of Oakland, 2017a
		U.S. Department of Transportation, 2012
		New York City Government, 2013
		The City and County of San Francisco, 2017a
		City of Philadelphia, 2017a
		City of Fort Worth, 2017a
		Blessing, 2015
	Vietnam	Japan International Cooperation Agency, 2004

Variable	Country	Sources
Rail Network (miles) Subway	Afghanistan	N/A
	Argentina	Buenos Aires City, 2010
	Bangladesh	N/A
	Brazil	Biderman, 2008
		CARIOCA, 2017
		Rodrigues & Silveira, 2016
	Canada	Toronto Transportation Commission, 2013
	China	China Data Online, 2017
	Democratic Republic of the Congo	N/A
	Egypt	Tour Egypt, 2017
	India	Indiastat, 2017 & Telangana Government, 2017
	Indonesia	MRT Jakarta
	Iran	N/A
	Iraq	Kneoma, 2017
	Kenya	N/A
	Mexico	Mexico City, 2009
	Nigeria	LAMATA, 2017
	Pakistan	N/A
	Peru	Railway Technology, 2017
	Philippines	N/A
	Russia	Moscow Metro, 2017
	Saudi Arabia	High Commission for the Development of ArRiyadh
		Varinsky, 2016
	South Africa	Gautrain Management Agency, 2016
	Thailand	Fernquest, 2016
		Railway Technology, 2017
	Turkey	TCDD, 2015 & Metro Istanbul, 2016
	United Kingdom	London Councils, 2017
	United States	Los Angeles County Metropolitan Transportation Authority, 2016
		Metropolitan Transit Authority, 2017
		Chicago Transit Authority, 2016
		Dallas Area Rapid Transit, 2017
		Bay Area Rapid Transit, 2017
	Vietnam	N/A

Variable	Country	Source
Sanitation Access (%)	Afghanistan	Kneoma, 2017
	Argentina	Engel, et al., 2011
	Bangladesh	World Bank, 2016
	Brazil	IBGE, 2010
	Canada	N/A
	China	National Bureau of Statistics of China, 2015
	Democratic Republic of the Congo	UN Data, 2017
	Egypt	Tour Egypt, 2017
	India	Global Water Forum, 2012
	Indonesia	World Bank, 2008
	Iran	N/A
	Iraq	Kneoma, 2017
	Kenya	Open Data for Africa, 2017d
	Mexico	UNICEF, 2013
	Nigeria	Kunnuji, 2014
	Pakistan	N/A
	Peru	National Institute of Statistics and Information, 2015
	Philippines	World Bank, 2015
	Russia	Kneoma, 2017
	Saudi Arabia	Tortajada, et al., 2006
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	N/A
	United Kingdom	N/A
	United States	World Bank, 2017c
	Vietnam	Van Leeuwen, Nyugen, & Dieperink, 2015

Variable	Country	Source
Hospitals	Afghanistan	Kneoma, 2017
	Argentina	Buenos Aires City, 2017
	Bangladesh	Google Maps Query
	Brazil	Google Maps Query
	Canada	Toronto, 2017a
	China	China Data Online, 2017
	Democratic Republic of the Congo	Google Maps Query
	Egypt	Tour Egypt, 2017
	India	Indiastat, 2017
	Indonesia	Jakarta Government, 2017
	Iran	Kneoma, 2017
	Iraq	Kneoma, 2017
	Kenya	Nairobi Governemt, 2017a
	Mexico	Google Maps Query
	Nigeria	Lagos State Government, 2013
	Pakistan	Kneoma, 2017
	Peru	Google Maps Query
	Philippines	Google Maps Query
	Russia	Kneoma, 2017
	Saudi Arabia	Kingdom of Saudi Arabia, 2017
	South Africa	City of Johannesburg, 2017
	Thailand	Google Maps Query
	Turkey	Istanbul Government, 2017
	United Kingdom	UK Health Centre, 2017
	United States	California Government, 2017
		Illinois Department of Public Health, 2016
		Indiana State Department of Health, 2017
		Maryland Government, 2017 & MHA, 2017
		New Hampshire Hospital Association,2017
		New York State, 2017
		PA Department of Health, 2013
		State of New Jersey, 2017
		Texas DHHS, 2012
		Washington, DC, 2017b
		Wisconsin Department of Health Services, 2017
	Vietnam	Ho Chi Minh City, 2005b

Variable	Country	Sources
Police Stations	Afghanistan	Google Maps Query
	Argentina	Ministry of Justice and Security,2017
	Bangladesh	
	Brazil	Google Maps Query
	Canada	Toronto Police, 2017
	China	Google Maps Query
	Democratic Republic of the Congo	Google Maps Query
	Egypt	Google Maps Query
	India	Google Maps Query
	Indonesia	Google Maps Query
	Iran	Google Maps Query
	Iraq	Google Maps Query
	Kenya	Nairobi Government, 2017a
	Mexico	Google Maps Query
	Nigeria	Open Data for Africa, 2017a
	Pakistan	Google Maps Query
	Peru	Google Maps Query
	Philippines	Google Maps Query
	Saudi Arabia	Google Maps Query
	South Africa	Google Maps Query
	Thailand	Google Maps Query
	Turkey	Istanbul Government, 2017
	United Kingdom	City of London Police, 2014

Variable	Country	Sources
Police Stations	United States	Anaheim Police, 2017
		Arlington, 2017
		Boston Police Department, 2017
		Calvert County, 2017
		Charles County 2017
		Chicago Police, 2017
		City of Alexandria, 2017
		City of Gary, Indiana, 2017
		City of Leesburg, 2017
		City of Oakland, 2017b
		Dallas Police Department, 2017
		Fairfax County, 2017
		Fort Worth Police, 2017
		Jersey City Police Department, 2017
		LAPD, 2017
		Long Beach, 2017
		New York City Government, 2017
		Newark Police Department, 2017
		Orange County, 2017
		Philadelphia Police, 2017
		Stafford County Sheriff, 2017
		The City and County of San Francisco, 2017b
		The City of Kenosha, 2017
		They City of Falls Church, 2017
		Washington, DC, 2017a
	Vietnam	Google Maps Query

Variable	Country	Sources
Fire Stations	Afghanistan	Google Maps Query
	Argentina	Google Maps Query
	Bangladesh	Google Maps Query
	Brazil	Google Maps Query
	Canada	Toronto, 2017b
	China	Google Maps Query
	Democratic Republic of the Congo	Google Maps Query
	Egypt	Google Maps Query
	India	Google Maps Query
	Indonesia	Jakarta Government, 2017
	Iran	Google Maps Query
	Iraq	Google Maps Query
	Kenya	Nairobi Governemt, 2017b
	Mexico	Google Maps Query
	Nigeria	Lagos State Government, 2017
	Pakistan	Google Maps Query
	Peru	Google Maps Query
	Philippines	Google Maps Query
	Russia	Google Maps Query
	Saudi Arabia	Google Maps Query
	South Africa	City of Johannesburg, 2017
	Thailand	Google Maps Query
	Turkey	Istanbul Fire Department, 2017
	United Kingdom	London Fire Brigade, 2017
	United States	United States Fire Administration
		New York City Fire Department, 2014
		Newark Fire Department, 2017
		Jersey City, 2017
	Vietnam	Google Maps Query

Variable	Country	Sources
Military Bases	Afghanistan	Google Maps Query
	Argentina	Google Maps Query
	Bangladesh	Google Maps Query
	Brazil	N/A
	Canada	N/A
	China	GlobalSecurity.org, 2017
	Democratic Republic of the Congo	Google Maps Query
	Egypt	Google Maps Query
	India	Google Maps Query
	Indonesia	Google Maps Query
	Iran	Google Maps Query
	Iraq	Google Maps Query
	Kenya	Google Maps Query
	Mexico	Google Maps Query
	Nigeria	Google Maps Query
	Pakistan	Google Maps Query
	Peru	Google Maps Query
	Philippines	Google Maps Query
	Russia	Google Maps Query
	Saudi Arabia	Google Maps Query
	South Africa	Google Maps Query
	Thailand	Google Maps Query
	Turkey	N/A
	United Kingdom	Google Maps Query
	United States	Department of Defense, 2017
	Vietnam	Google Maps Query

Variable	Country	Source
State/Provincial Capital National Capital	Afghanistan	Kneoma, 2017
	Argentina	Kneoma, 2017
	Bangladesh	Kneoma, 2017
	Brazil	Kneoma, 2017
	Brazil	Kneoma, 2017
	Canada	Kneoma, 2017
	China	Kneoma, 2017
	Democratic Republic of the Congo	Kneoma, 2017
	Egypt	Kneoma, 2017
	India	Kneoma, 2017
	Indonesia	Kneoma, 2017
	Iran	Kneoma, 2017
	Iraq	Kneoma, 2017
	Kenya	Kneoma, 2017
	Mexico	Kneoma, 2017
	Nigeria	Kneoma, 2017
	Pakistan	Kneoma, 2017
	Peru	Kneoma, 2017
	Philippines	Kneoma, 2017
	Russia	Kneoma, 2017
	Saudi Arabia	Kneoma, 2017
	South Africa	Kneoma, 2017
	Thailand	Kneoma, 2017
	Turkey	Kneoma, 2017
	United Kingdom	Kneoma, 2017
	United States	Kneoma, 2017
	Vietnam	Kneoma, 2017

Variable	Country	Sources
GDP (billion USD)	Afghanistan	Kneoma, 2017
	Argentina	Buenos Aires City, 2010
	Bangladesh	The University of Tokyo
	Brazil	IBGE, 2010
	Canada	Toronto, 2016
	China	China Data Online, 2017
	Democratic Republic of the Congo	UN Data, 2017
	Egypt	Open Data for Africa, 2017b
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	Kneoma, 2017
	Iraq	Kneoma, 2017
	Kenya	Open Data for Africa, 2017d
	Mexico	Kneoma, 2017
	Mexico	Tortajada, et al., 2006
	Nigeria	Open Data for Africa, 2017a
	Pakistan	Lloyd's, 2014
	Peru	National Institute of Statistics and Information, 2015
	Philippines	Kneoma, 2017
	Russia	Kneoma, 2017
	Saudi Arabia	Kingdom of Saudi Arabia, 2017
	South Africa	Open Data for Africa, 2017a
	Thailand	Kneoma, 2017
	Turkey	Turkey Statistical Institute, 2017a
	United Kingdom	Eurostat, 2017
	United States	United States Department of Commerce, 2015
	Vietnam	Statistical Documentation and Service Center, 2017
	Vietnam	Ho Chi Minh City, 2005a

Variable	Country	Source
High School Diploma (%) Bachelor's Degree (%)	Afghanistan	N/A
	Argentina	Buenos Aires City, 2010
	Bangladesh	Kneoma, 2017
	Brazil	Cidades, 2010a
	Brazil	Cidades, 2010b
	Canada	Statistics Canada, 2011
	China	China Data Online, 2017
	Democratic Republic of the Congo	N/A
	Egypt	N/A
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	N/A
	Iraq	N/A
	Kenya	USAID, 2003
	Mexico	INEGI, 2017
	Nigeria	N/A
	Pakistan	N/A
	Peru	National Institute of Statistics and Information, 2015
	Philippines	N/A
	Russia	Mosgorstat, 2010
	Saudi Arabia	N/A
	Siuth Africa	N/A
	Thailand	N/A
	Turkey	Kneoma, 2017
	United Kingdom	N/A
	United States	United States Census Bureau, 2015a
	Vietnam	N/A

Variable	Country	Sources
Literacy Rate (%)	Afghanistan	Kneoma, 2017
	Argentina	Buenos Aires City, 2010
	Bangladesh	Kneoma, 2017
	Brazil	IBGE, 2010
	Canada	N/A
	China	China Data Online, 2017
	Democratic Republic of the Congo	N/A
	Egypt	Tour Egypt, 2017
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	Kneoma, 2017
	Iraq	Kneoma, 2017
	Kenya	N/A
	Mexico	UNICEF, 2013
	Nigeria	Open Data for Africa, 2017a
	Pakistan	Kneoma, 2017
	Peru	National Institute of Statistics and Information, 2015
	Philippines	Kneoma, 2017
	Russia	N/A
	Saudi Arabia	N/A
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	N/A
	United Kingdom	Kneoma, 2017
	United States	Literacy Partners, 2017
		Los Angeles Almanac, 2017
		Mafrica, L., 2009
		Texas Department of Health and Human Services, 2003
		Literacy Coalition, 2017
		Reaves, 2010
		Alexander, 2007
	Vietnam	Statistical Documentation and Service Center, 2017

Variable	Country	Source
Average Income (USD)	Afghanistan	The Asia Foundation, 2015
	Argentina	Buenos Aires City, 2010
	Bangladesh	The University of Tokyo
	Brazil	IBGE, 2010
	Canada	Statistics Canada, 2011
	China	China Data Online, 2017
	Democratic Republic of the Congo	N/A
	Egypt,	Survey Explorer, 2017
	India	Indiastat, 2017
	Indonesia	Jakarta Government, 2017
	Iran	Homylafayette, 2011
	Iraq	N/A
	Kenya	N/A
	Mexico	UNICEF, 2013
	Nigeria	N/A
	Pakistan	N/A
	Peru	National Institute of Statistics and Information, 2015
	Philippines	Kneoma, 2017
	Russia	Kneoma, 2017
	Saudi Arabia	N/A
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	Turkey Statistical Institute, 2017a
	United kingdom	United Kingdom, 2017
	United States	United States Census Bureau, 2015a
	Vietnam	Nam, 2016

Variable	Country	Source
Poverty Rate (%)	Afghanistan	Kneoma, 2017
	Argentina	N/A
	Bangladesh	Sohel, 2014
	Brazil	Cidades, 2010a
	Brazil	Cidades, 2010b
	Canada	Toronto, 2013
	Canada	Government of Canada, 2017
	China	China Data Online, 2017
	Democratic Republic of the Congo	Open Data for Africa, 2017e
	Egypt	Open Data for Africa, 2017b
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	Homylafayette, 2011
	Iraq	Kneoma, 2017
	Kenya	Open Data for Africa, 2017d
	Mexico	INEGI, 2017
	Nigeria	Open Data for Africa, 2017a
	Pakistan	News Reports, 2013
	Peru	National Institute of Statistics and Information, 2015
	Philippines	Kneoma, 2017
	Russia	Zykov, 2015
	Saudi Arabia	N/A
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	Turkey Statistical Institute, 2017a
	United Kingdom	Trust for London and New Policy Institute, 2015
	United States	United States Census Bureau, 2015a
	Vietnam	Tuoitre News, 2014

Variable	Country	Source
Average Household Size	Afghanistan	Kneoma, 2017
	Argentina	Buenos Aires City, 2010
	Bangladesh	Kneoma, 2017
	Brazil	IBGE, 2010
	Canada	Statistics Canada, 2011
	China	China Data Online, 2017
	Democratic Republic of the Congo	Open Data for Africa, 2017e
	Egypt	Tour Egypt, 2017
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	Kneoma, 2017
	Iraq	Kneoma, 2017
	Kenya	USAID, 2003
	Mexico	Kneoma, 2017
	Nigeria	Open Data for Africa, 2017a
	Pakistan	Kneoma, 2017
	Peru	N/A
	Philippines	Kneoma, 2017
	Russia	Mosgorstat, 2010
	Saudi Arabia	Kneoma, 2017
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	Turkey Statistical Institute, 2013c
	Turkey	Turkey Statistical Institute, 2017a
	United Kingdom	Nomis, 2011
	United States	United States Census Bureau, 2015a
	Vietnam	Ministry of Planning and Investment, 2011b

Variable	Country	Source
Pop Under 18–19 Yrs Old (%) Pop Over 65 Yrs Old (%)	Afghanistan	Kneoma, 2017
	Argentina	Buenos Aires City, 2010
	Bangladesh	N/A
	Brazil	IBGE, 2010
	Canada	Statistics Canada, 2011
	China	China Data Online, 2017
	Democratic Republic of the Congo	Open Data for Africa,2017e
	Egypt	Open Data for Africa, 2017b
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	Kneoma, 2017
	Iraq	Kneoma, 2017
	Kenya	N/A
	Mexico	INEGI, 2017
	Nigeria	Lagos State Government, 2013
	Pakistan	Kneoma, 2017
	Peru	National Institute of Statistics and Information, 2015
	Philippines	Kneoma, 2017
	Russia	Mosgorstat, 2010
	Saudi Arabia	Kingdom of Saudi Arabia, 2017
	South Africa	Open Data for Africa, 2017c
	Thailand	N/A
	Turkey	Turkey Statistical Institute, 2017a
	United Kingdom	Nomis, 2011
	United States	United States Census Bureau, 2015a
	Vietnam	Ministry of Planning and Investment, 2011a

Variable	Country	Source
Internet Access (%)	Afghanistan	Kneoma, 2017
	Argentina	Buenos Aires City, 2010
	Bangladesh	Kneoma, 2017
	Brazil	IBGE, 2010
	Canada	N/A
	China	CINIC, 2017
	Democratic Republic of the Congo	UN Data, 2017
	Egypt	N/A
	India	
	Indonesia	BPS, 2017
	Iran	N/A
	Iraq	Kneoma, 2017
	Kenya	N/A
	Mexico	INEGI, 2017
	Nigeria	N/A
	Pakistan	National Institute of Population Studies, 2013
	Peru	National Institute of Statistics and Information, 2015
	Philippines	Kneoma, 2017
	Russia	Yandex, 2016
	Saudi Arabia	Kingdom of Saudi Arabia, 2017
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	N/A
	United Kingdom	Kneoma, 2017
	United States	United States Census Bureau, 2015b
	Vietnam	Cimigo, 2011

Variable	Country	Source
Pop Density (per sq. mile)	Afghanistan	Kneoma, 2017
	Argentina	Engel, et al., 2011
	Bangladesh	Kneoma, 2017
	Brazil	Cidades, 2010a
	Brazil	Cidades, 2010b
	Canada	Statistics Canada, 2011
	China	China Data Online, 2017
	Democratic Republic of the Congo	Open Data for Africa, 2017a
	Egypt	Tour Egypt, 2017
	India	Indiastat, 2017
	Indonesia	BPS, 2017
	Iran	Kneoma, 2017
	Iraq	Kneoma, 2017
	Kenya	Open Data for Africa, 2017d
	Mexico	Kneoma, 2017
	Nigeria	Open Data for Africa, 2017a
	Pakistan	United Nations, 2015
	Peru	National Institute of Statistics and Information, 2015
	Philippines	Kneoma, 2017
	Russia	Kneoma, 2017
	Saudi Arabia	Kneoma, 2017
	South Africa	City of Johannesburg, 2017
	Thailand	Kneoma, 2017
	Turkey	Kneoma, 2017
	United Kingdom	London Councils, 2017
	United States	United States Census Bureau, 2015
	Vietnam	Statistical Documentation and Service Center, 2017

Variable	**Country**	**Source**
Primary Religion	Afghanistan	Central Intelligence Agency, 2017
	Argentina	Central Intelligence Agency, 2017
	Bangladesh	Central Intelligence Agency, 2017
	Brazil	Central Intelligence Agency, 2017
	Brazil	Central Intelligence Agency, 2017
	Canada	Central Intelligence Agency, 2017
	China	Central Intelligence Agency, 2017
	Democratic Republic of the Congo	Central Intelligence Agency, 2017
	Egypt	Central Intelligence Agency, 2017
	India	Central Intelligence Agency, 2017
	Indonesia	Central Intelligence Agency, 2017
	Iran	Central Intelligence Agency, 2017
	Iraq	Central Intelligence Agency, 2017
	Kenya	Central Intelligence Agency, 2017
	Mexico	Central Intelligence Agency, 2017
	Nigeria	Central Intelligence Agency, 2017
	Pakistan	Central Intelligence Agency, 2017
	Peru	Central Intelligence Agency, 2017
	Philippines	Central Intelligence Agency, 2017
	Russia	Central Intelligence Agency, 2017
	Saudi Arabia	Central Intelligence Agency, 2017
	South Africa	Central Intelligence Agency, 2017
	Thailand	Central Intelligence Agency, 2017
	Turkey	Central Intelligence Agency, 2017
	United Kingdom	Central Intelligence Agency, 2017
	United States	Central Intelligence Agency, 2017
	Vietnam	Central Intelligence Agency, 2017

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Alexander, K. L. (2007, March 19). Illiteracy aid found to lag in district. *The Washington Post*. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2007/03/18/AR2007031801347.html

Anaheim Police. (2017). Locations. Retrieved from http://www.anaheim.net/363/Locations

Aplin, S. (2013, February 26). Nuclear powers Toronto, cheaply and with no carbon. Retrieved from https://canadianenergyissues.com/2013/02/26/nuclear-powers-toronto-cheaply-and-with-no-carbon/

Arlington County. (2015). *2015 Annual water quality report*. Retrieved from https://arlingtonva.s3.dualstack.us-east-1.amazonaws.com/wp-content/uploads/sites/4/2013/09/ArlingtonWaterQualityReport2015.pdf

Arlington. (2017). Police districts. Retrieved from https://police.arlingtonva.us/about/police-districts/

ASEAN Ports. (2001). Bangkok Port. Retrieved from https://web.archive.org/web/20080930103827/http://www.aseanports.com/apa_members/apa_thai/apa_thai1.htm

Asia Foundation. (2015). *A survey of the Afghan people*. Retrieved from https://asiafoundation.org/resources/pdfs/Afghanistanin2015.pdf

Barbooti, M. M., Bolzoni, G., Mirza, I. A., Pelosi, M., Barilli, L., Kadhum, R., & Peterlongon, G. (2010). Evaluation of quality of drinking water from Baghdad, Iraq. *Science World Journal,* 10(2), 35–46. Retrieved from https://www.ajol.info/index.php/swj/article/viewFile/61512/49643

Barnett, T. (2004). The Pentagon’s new map: War and peace in the twenty-first century. New York, NY: G.P. Putnam’s Sons.

Bay Area Rapid Transit. (2017). System facts. Retrieved from https://www.bart.gov/about/history/facts

BBC News. (2017, Jan. 11). Ukraine power cut was a cyber-attack. Retrieved from http://www.bbc.com/news/technology-38573074

BCPT. (2017). *Mumbai’s water supply*. Retrieved from http://www.bcpt.org.in/articles/watersupply.pdf

Biderman, C. (2008, December). São Paulo’s urban transport infrastructutre. Retrieved from https://lsecities.net/media/objects/articles/sao-paulo-urban-transport-infrastructure/en-gb/

Blessing, K. (2014, Sept 5). City workers will walk Boston streets, offer their views. Retrieved from https://www.bostonglobe.com/metro/2014/09/05/boston-workers-walk-all-miles-city-streets/ZNJY2BqHm5mz1f9XMcdcuM/story.html

Boston Police Department. (2017). Districts. Retrieved from http://bpdnews.com/districts/

BPS. (2017). Statistics Indonesia. Retrieved from https://www.bps.go.id/index.php

BTRC. (2015, Apr.). Number of mobile phone subscribers. Retrieved from http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-april-2015

Buenos Aires City. (2010). Statistics and censuses: Data bank. Retrieved from http://www.estadisticaciudad.gob.ar/eyc/?cat=111

Buenos Aires City. (2017). Establishments - hospitals and health centers. Retrieved from Health: http://www.buenosaires.gob.ar/salud/establecimientos

Buenos Aires Port. (2016). *Buenos Aires port statistics 2016*. Retrieved from http://www.puertobuenosaires.gov.ar/ver_archivos/estadisticas-2016/108.pdf

BWSC. (2017). Boston Water and Sewer Commission: Water sources. Retrieved from http://www.bwsc.org/ABOUT_BWSC/systems/water/sources.asp

BWSSB. (2016). Bangalore Water Supply and Sewerage Board: About BWSSB. Retrieved from https://bwssb.gov.in/content/about-bwssb-2

CAA, Bangladesh. (2006). Airports. Retrieved from http://www.caab.gov.bd/adinfo/airports.html

Cairo Electricity Production Company. (2017). Stations. Retrieved from http://www.cairoepc.com/EN/cnorth.html

California Government. (2016). Hospitals. Retrieved from http://www.oshpd.ca.gov/HID/Facility-Listing.html

Calvert County. (2017). Calvert County Sheriff’s Office: Office locations. Retrieved from http://www.co.cal.md.us/index.aspx?nid=384

Campbell, J. (2015). A dry and ravaged land: Investigating water resources in Afghanistan. Retrieved from https://www.earthmagazine.org/article/dry-and-ravaged-land-investigating-water-resources-afghanistan

CARIOCA. (2017). About VLT. Retrieved from http://vltrio.rio/o-projeto/

Central Intelligence Agency. (2017). World factbook. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/fields/2177.html

Charles County. (2017). Charles Country Sheriff's Office: Installations. Retrieved from https://www.ccso.us/our-installations/

Chatterjee, A. (2014, Nov. 24). Water supply system in Kolkata and ajoining areas. Retrieved from https://medium.com/@anjan.chatterjee/water-supply-system-in-kolkata-city-and-adjoining-areas-b199099a4517

Chicago Police. (2017). Districts. Retrieved from http://home.chicagopolice.org/community/districts/

Chicago Transit Authority. (2016). CTA facts at a glance. Retrieved from http://www.transitchicago.com/about/facts.aspx

China Data Online. (2017). City statistics. Retrieved from http://chinadataonline.org/

Cidades. (2010a). São Paulo. Retrieved from : http://cidades.ibge.gov.br/xtras/perfil.php?codmun=355030&lang=_EN

Cidades. (2010b). Rio de Janeiro. Retrieved from http://cidades.ibge.gov.br/xtras/temas.php?lang=_EN&codmun=330455&idtema=105&search=rio-de-janeiro|rio-de-janeiro|2010-population-census:-results-of-the-sample-education-

Cimigo. (2011). *2011 Vietnam NetCitizens report*. Retrieved from http://www.cimigo.com/en/download/research_report/348.

CINIC. (2017). *China statistical report on internet development*. Retrieved from http://www.cnnic.com.cn/hlwfzyj/hlwxzbg/hlwtjbg/201701/P020170123364672657408.pdf

City and County of San Francisco. (2017a). Miles of streets. Retrieved from https://data.sfgov.org/City-Infrastructure/Miles-Of-Streets/5s76-j52p/data

City and County of San Francisco. (2017b). Police district Maps. Retrieved from http://sanfranciscopolice.org/police-district-maps

City of Alexandria. (2017). Police department: Contact information. Retrieved from https://www.alexandriava.gov/police/info/default.aspx?id=902

City of Chicago. (2017a). Streets, alleys, & sidewalks. Retrieved from https://www.cityofchicago.org/city/en/depts/cdot/provdrs/street.html

City of Chicago. (2017b). Water management. Retrieved from https://www.cityofchicago.org/city/en/depts/water/provdrs/supply.html

City of Dallas. (2017). Water quality information. Retrieved from http://dallascityhall.com/departments/waterutilities/Pages/water_quality_information.aspx

City of Falls Church. (2017). Divisions. Retrieved from http://www.fallschurchva.gov/295/Divisions

City of Fort Worth. (2017a). Transportation & public works. Retrieved from http://fortworthtexas.gov/tpw/

City of Fort Worth. (2017b). Water supply. Retrieved from http://fortworthtexas.gov/water/drinking-water/supply/

City of Gary. (2017). Police department. Retrieved from http://www.gary.in.us/city-departments/police-department.asp

City of Johannesburg. (2017). *Statistical publications: Issues 1 and issues 2*. Retrieved from https://joburg.org.za/index.php?option=com_content&view=article&id=11317%25252525252525253Acity-governance&catid=84&Itemid=131

City of Kenosha. (2017). Kenosha police. Retrieved from https://www.kenosha.org/departments/police/

City of Leesburg. (2017). Contact Leesburg police. Retrieved from http://www.leesburgva.gov/government/departments/police-department/contact-leesburg-police

City of London Police. (2014). Police stations. Retrieved from https://www.cityoflondon.police.uk/contact-city-police/police-stations/Pages/default.aspx

City of Los Angeles. (2008). State of the streets. Retrieved from http://bss.lacity.org/State_Streets/StateOfTheStreets.htm

City of Newark. (2015). 2014 *Annual water report*. Retrieved from https://ndex.ci.newark.nj.us/dsweb/Get/Document-525123/2014%20Annual%20Water%20Quality%20Report%20CCR.pdf

City of Oakland. (2017). Oakland streets fact sheet. Retrieved from www2.oaklandnet.com/w/oak029773

City of Oakland. (2017). Police department. Retrieved from http://www2.oaklandnet.com/government/o/OPD/a/contact/index.htm

City of Philadelphia. (2017a). About the streets dept. & its division. Retrieved from http://www.philadelphiastreets.com/about/

City of Philadelphia. (2017b). Water utility. Retrieved from http://www.phila.gov/WATER/WU/Pages/default.aspx

Dallas Area Rapid Transit. (2017). Frequently asked questions about DART. Retrieved from http://www.dart.org/aptarraildallas/DARTFAQ.pdf

Dallas Police Department. (2017). Divisions. Retrieved from http://www.dallaspolice.net/division

DC Water. (2017). Home. Retrieved from https://www.dcwater.com/

Delhi Government. (2015). Water supply and sewerage. Retrieved from http://www.delhi.gov.in/wps/wcm/connect/16541d8048d8ebdea8e9f97a2b587979/ESD+2014-15+-+Ch-13.pdf?MOD=AJPERES&lmod=519320373&CACHEID=16541d8048d8ebdea8e9f97a2b587979

Demirci, A., & Butt, A. (2001). Historical overview and current trends in Istanbul water supply development. *Globalization and Water Resources Management: The Changing Value of Water*. Retrieved from http://www.awra.org/proceedings/dundee01/Documents/DemirciandButtfinal.pdf

Demographia. (2016). *Demographia world urban areas: 12th edition*. Retrieved from http://www.demographia.com/db-worldua.pdf

Department for Business, Energy & Industrial Strategy. (2016). Sub-national electricity and gas consumption statistics. Retrieved May 23, 2017, from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/579203/Sub-national_electricity_and_gas_consumption_summary_report_2016.pdf

Department of Defense. (2017). Retrieved from Military Installations: http://www.militaryinstallations.dod.mil/

DHA Cogen LTD. (2017). Welcome to DHA Cogen LTD. Retrieved from http://www.dhacogen.com/

Dickenson, M. (2014). *Solving operational models of interdependent infrastructure systems*. (Master’s thesis), Retrieved from http://calhoun.nps.edu/bitstream/handle/10945/44550/14Dec_Dickenson_Michael.pdf?sequence=1

Do, T. T. (2002). Ho Chi Minh City: The source and challenge for the next year. *Proceedings of IGES/APN Megacities Project, IGES*. Retrieved from https://enviroscope.iges.or.jp/contents/13/data/PDF/09-2HO%20CHI%20MINH%20CITY%20ENERGY.pdf

Dobbs, R., Smit, S., Remes, J., Manyika, J., Roxburgh, C., & Restrepo, A. (2011). *Urban world: Mapping the economic power of cities*. Retrieved from http://www.mckinsey.com/global-themes/urbanization/urban-world-mapping-the-economic-power-of-cities

DPWH. (2016). National road length by surface type. Retrieved from http://www.dpwh.gov.ph/dpwh/2016%20DPWH%20ATLAS/Tables%20&%20Graphs%20(Roads)/Road%20Data%202016/ATLAS%202016/Table%201.1c.htm

East Bay Municipal Utility District. (2017). *Water supply engineering daily report*. Retrieved from http://www.ebmud.com/water-and-drought/about-your-water/water-supply/water-supply-reports/daily-water-supply-report/

Engel, K., Jokiel, D., Kraljevic, A., Geiger, M., & Smith, K. (2011). *Big cities. Big water. Big challenges. Water in an urbanizing world*. Retrieved from http://www.wwf.se/source.php/1390895/Big%20Cities_Big%20Water_Big%20Challenges_2011.pdf

Eskom. (2017). Power stations. Retrieved from http://www.eskom.co.za/Whatweredoing/ElectricityGeneration/PowerStations/Pages/Power_Stations_And_Pumped_Storage_Schemes.aspx

Eurostat. (2017). Database. Retrieved from http://ec.europa.eu/eurostat/data/database

Fairfax County. (2017). Police stations. Retrieved from http://www.fairfaxcounty.gov/police/stations/

Fairfax Water. (2017). Treatment. Retrieved from http://www.fcwa.org/education/treatment/WT-web-final.html

Felix, K., & Wong, F. (2015). The case for megacities. *U.S. Army War College Quarterly Parameters*, *19–32*. Retrieved from https://ssi.armywarcollege.edu/pubs/parameters/Issues/Spring_2015/5_FelixKevin_WongFrederick_The%20Case%20for%20Megacities.pdf

Fernquest, J. (2016, Aug. 22). Purple Line: Newly launched but few using it. Retrieved from http://www.bangkokpost.com/learning/advanced/1068305/purple-line-newly-launched-but-few-using-it

Fort Worth Police. (2017). Patrol division. Retrieved from https://www.fortworthpd.com/Divisions/Patrol.aspx

Gautrain Management Agency. (2016). Socio-economic development process. Retrieved from http://www.gautrain.co.za/contents/brochures/sed_brochure_final_print.pdf

Global Water Forum. (2012). Water supply and sanitation in India: Meeting targets and beyond. Retrieved from http://www.globalwaterforum.org/2012/09/23/water-supply-and-sanitation-in-india-meeting-targets-and-beyond/

GlobalSecurity.org. (2017). China military guide. Retrieved from http://www.globalsecurity.org/military/world/china/index.html

Google Maps Query (2017). Google. Retrieved from https://www.google.com/maps

Government of Canada. (2017). Population of census metropolitan areas. Retrieved from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo05a-eng.htm

Government of Mexico. (2017). Effective capacity by federative entity of electrical energy. Retrieved from https://datos.gob.mx/busca/dataset/capacidad-efectiva-por-entidad-federativa-de-energia-electrica

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika, 325–328*. Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cmdscale.html

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27, 857–871*. Retrieved from https://www.jstor.org/stable/2528823?seq=1#page_scan_tab_contents

Greater London Authority. (2017). Climate change, weather and water. Retrieved from https://www.london.gov.uk/what-we-do/environment/climate-change-weather-and-water

Hanushek, E., & Woessmann, L. (2010). Education and economic growth. *International Encyclopedia of Education, 245–252*. Retrieved from http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BWoessmann%202010%20IntEncEduc%202.pdf

Harris, T., & Ross, F. (1955). Fundamentals of a method for evaluating rail net capacities. RAND Corporation. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/093458.pdf

High Commission for the Development of ArRiyadh. (2017). Transport. Retrieved from http://www.arriyadh.com/Eng/Ab-Arriyad/Left/Geography/getdocument.aspx?f=/openshare/Eng/Ab-Arriyad/Left/Geography/Transport.doc_cvt.htm

HMWSSB. (2015). Hyderabad Metropolitan Water Supply and Sewerage Board: Source. Retrieved from https://www.hyderabadwater.gov.in/en/index.php/about/source

Ho Chi Minh City. (2005a). Economic hub. Retrieved from http://www.eng.hochiminhcity.gov.vn/abouthcmcity/Lists/Posts/Post.aspx?CategoryId=13&ItemID=5509&PublishedDate=2005-03-06T10:43:42Z

Ho Chi Minh City. (2005b). Core data. Retrieved from http://www.eng.hochiminhcity.gov.vn/abouthcmcity/Lists/Posts/Post.aspx?CategoryId=15&ItemID=5524&PublishedDate=2005-03-08T11:26:49Z

Homylafayette. (2011, Mar. 4). Iran’s cities a sea of poverty. Retrieved from http://www.pbs.org/wgbh/pages/frontline/tehranbureau/2011/03/irans-cities-a-sea-of-poverty.html

Hornik, K. (2005). A clue for cluster ensembles. *Journal of Statistical Software, 14(12)*. Retrieved from https://www.jstatsoft.org/article/view/v014i12

Hurriyet Daily News. (2014, Mar. 17). Electricity consumption in Turkey soars 78 pct over last decade. Retrieved from http://www.hurriyetdailynews.com/electricity-consumption-in-turkey-soars-78-pct-over-decade.aspx?pageID=238&nID=63699&NewsCatID=348

Hussain. (2011, Dec. 24). In Karachi, 16,562 more vehicles hit the roads each month. Retrieved from https://www.pakistantoday.com.pk/2011/12/24/in-karachi-16562-more-vehicles-hit-the-roads-each-month/

IAPH. (2015). World container traffic data 2015. Retrieved from http://www.iaphworldports.org/iaph/wp-content/uploads/WorldPortTraffic-Data-for-IAPH-using-LL-data2015.pdf

IBGE. (2010). Municipal social indicators: an analysis of the results of the demographic census. Retrieved from http://www.ibge.gov.br/home/estatistica/populacao/censo2010/indicadores_sociais_municipais/indicadores_sociais_municipais_tab_uf_zip.shtm

ICS-CERT. (2016, Feb. 25). Cyber-attack against Ukrainian infrastructure. Retrieved from https://ics-cert.us-cert.gov/alerts/IR-ALERT-H-16-056-01

Illinois Department of Public Health. (2017). IDPH hospital directory. Retrieved from https://data.illinois.gov/Public-Health/IDPH-Hospital-Directory/wsms-teqm

Indiana State Department of Health. (2017). Indiana hospital directory. Retrieved from http://www.in.gov/isdh/reports/QAMIS/hosdir/

Indiastat. (2017). Socio-economic statistical information about India. Retrieved from indiastat.com

INEGI. (2017). Statistics: Themes. Retrieved from http://www.inegi.org.mx/

Institute for Global Environmental Studies. (2007). *Sustainable groundwater management in Asian cities*. Freshwater Resources Management Project. Retrieved from https://pub.iges.or.jp/system/files/publication_documents/pub/policyreport/654/00_complete_report.pdf

Istanbul Fire Department. (2017). Our stations. Retrieved from http://itfaiye.ibb.gov.tr/tr/istasyonlarimiz.html

Istanbul Government. (2017). Info Istanbul. Retrieved from http://istanbul.gov.tr/en/info-istanbul/

Jakarta Government. (2014). *Jakarta urban transportation problems and their environmental impact*. Retrieved from http://www.ui.ac.id/download/apru-awi/jakarta-local-goverment.pdf

Jakarta Government. (2017). Jakarta open data. Retrieved from http://data.jakarta.go.id/

Japan International Cooperation Agency. (2004). *The study on the urban transport master plan and feasibility study in Ho Chi Minh City*. Retrieved from http://open_jicareport.jica.go.jp/710/710/710_123_11764636.html

Jersey City Police Department. (2017). Organization. Retrieved from http://www.njjcpd.org/organization

Jersey City. (2017). Jersey City Fire Department. Retrieved from http://www.cityofjerseycity.com/emergency.aspx?id=15062

Johannesburg Water. (2017). Understanding water. Retrieved from https://www.johannesburgwater.co.za/water-and-sanitation/understanding-water/

Karachi Port Trust. (2016). Cargo, container handling & shipping. Retrieved from http://kpt.gov.pk/pages/default.aspx?id=32

Karachi Water & Sewerage Board. (2013). Water management. Retrieved from http://www.kwsb.gos.pk/View.aspx?Page=25

Karanja, J. (2011). Improving water provision in Nairobi through control of non-revenue water. *Global water summit 2011.* Retrieved from https://www.globalwaterintel.com/client_media/uploaded/Conference%20book%202011_Focusing_on_performance_72%20pixels%20per%20inch.pdf

Kaufman, L., & Rousseuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons.

KGM. (2017). State and provincial road inventory. Retrieved from http://www.kgm.gov.tr/Sayfalar/KGM/SiteTr/Istatistikler/DevletveIlYolEnvanteri.aspx

Kilcullen, D. (2013). *Out of the mountains: The coming age of the urban guerrilla*. New York, NY. Oxford University Press.

Kingdom of Saudi Arabia. (2017). General authority for statistics. Retrieved from https://www.stats.gov.sa/en

Kneoma. (2017). World data atlas. Retrieved from https://knoema.com/atlas

Koop, F. (2015, May 9). Smartphones more prevalent in Argentina (but hard to find). Retrieved from http://www.buenosairesherald.com/article/188754/smartphones-more-prevalent-in-argentina-(but-hard-to-find)

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software, 74(7)*. Retrieved from https://www.jstatsoft.org/article/view/v074i07

Kunnuji, M. (2014). Spatial variations in access to improved sanitation and water in Lagos State. *Journal of Water, Sanitation, and Hygiene Development, 4(4), 612–619*. Retrieved from http://washdev.iwaponline.com/content/4/4/612

Lagos State Government. (2013). *2013 Digest of statistics*. Retrieved from http://lagosbudget.org/digest-of-statistics-2013

Lagos State Government. (2017). Lagos State Fire Service. Retrieved from http://fireservice.lagosstate.gov.ng/

LAMATA. (2017). Lagos rail mass transit. Retrieved from http://www.lamata-ng.com/rail_services.php

Landing and Planning. (2017). Sources of water supply Moscow. Retrieved from http://terraplan.ru/arhiv/21-1-3-2006/145-88.html

LAPD. (2017). Community police station address directory. Retrieved from http://www.lapdonline.org/our_communities/content_basic_view/6279

Lepeska, D. (2011, Nov. 9). Why $1B doesn’t buy much transit infrastructure anymore. Retrieved from https://www.citylab.com/transportation/2011/11/1-billion-doesnt-buy-much-transit-infrastructure-anymore/456/

Literacy Coalition. (2017). About the coalition. Retrieved from http://dallaslibrary2.org/literacy/coalition.php

Literacy Partners. (2017). The challenge. Retrieved from https://literacypartners.org/literacy-in-new-york-city-the-challenge

LIWA. (2014). Sustainable water and wastewater management in urban centres coping with climate change - concepts for metropolitan Lima. Retrieved from http://www.lima-water.de/en/lima.html

Lloyd's. (2014). Lloyd's city risk index 2015–2025. Retrieved May 22, 2017, from University of Cambridge: https://www.lloyds.com/cityriskindex/locations/fact_sheet/Karachi

London Councils. (2017). London facts and statistics. Retrieved from http://www.londoncouncils.gov.uk/who-runs-london/london-facts-and-statistics

London Fire Brigade. (2017). A-Z fire stations. Retrieved from http://www.london-fire.gov.uk/A-ZFireStations.asp

Long Beach. (2017). Police department phone list. Retrieved from http://www.longbeach.gov/police/contact-us/contact-us/

Los Angeles. (2017). Department of dater and power. Retrieved from https://www.ladwp.com/ladwp/faces/ladwp/aboutus/a-water/a-w-sourcesofsupply;jsessionid=sTCHZprL9KBJzb6wDymjlw8pz9G9XDM8kWbXrR9rnc8gzcGZgRTG!-248894535?_adf.ctrl-state=9zgjjuoqk_4&_afrLoop=215074901377344&_afrWindowMode=0&_afrWindowId=null#%40%3F_afrWindow

Los Angeles Almanac. (2017). Literacy in Los Angeles County. Retrieved from http://www.laalmanac.com/education/ed33a.php

Los Angeles County Metropolitan Transportation Authority. (2016). Bus & rail. Retrieved from Mtro: https://www.metro.net/news/facts-glance/

MacQueen, J. (1967). Some methods for classification and analysis of mulitvariate observations. *Proc. Fifth Berkeley Symposium on Math. Stats and Prob*. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512992

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2016). Cluster: Cluster analysis basics and extensions. R package version 2.0.4.

Mafrica, L. (2009, Jan. 20). Literacy rates challenge in Philadelphia. Retrieved from http://temple-news.com/opinion/literacy-rates-challenge-philadelphia/

Mahmud, A. H. (2014, Sept. 8). Dhaka City's infrastructure in a shambles. Retrieved from http://archive.dhakatribune.com/bangladesh/2014/sep/08/dhaka-city%E2%80%99s-infrastructure-shambles

MARDAS. (2012). Istanbul Ambarli Port. Retrieved from http://www.mardas.com.tr/acentelik/mardas.aspx?id=25&lang=en

Maryland Government. (2017). Hospitals. Retrieved from http://msa.maryland.gov/msa/mdmanual/01glance/html/hospital.html

Metro Istanbul. (2016). Metro Istanbul in brief. Retrieved from http://www.metro.istanbul/about-us/introduction.aspx

Metropolitan Electricity Authority. (2014). *Annual report 2014*. Retrieved from http://www.mea.or.th/en/e-magazine/detail/2786/254

Metropolitan Transportation Authority. (2017). Subways. Retrieved from http://web.mta.info/nyct/facts/ffsubway.htm

Mexico City. (2009). Transportation and roads in the federal district. Retrieved from https://web.archive.org/web/20090815115257/http://www.setravi.df.gob.mx/vialidades/transporte_vialidad.html

MHA. (2017). Massachusetts hospitals: Directory. Retrieved from http://www.mhalink.org/AM/Template.cfm?Section=Hospital_Directory

Ministry of Commerce. (2007). Doing business in China: Surveys. Retrieved from https://archive.is/20130805091244/http://english.mofcom.gov.cn/article/zt_business/lanmub/

Ministry of Energy and Mines. (2015). *Executive directory of electricity 2014*. Retrieved from http://www.minem.gob.pe/_publicaSector.php?idSector=6

Ministry of Justice and Security. (2017). City police. Retrieved from http://www.policiadelaciudad.gob.ar/?q=comisarias

Ministry of Mines and Energy. (2015). *2015 Statistical yearbook of electricity*. Retrieved from http://www.epe.gov.br/AnuarioEstatisticodeEnergiaEletrica/Anu%C3%A1rio%20Estat%C3%ADstico%20de%20Energia%20El%C3%A9trica%202015.pdf

Ministry of Planning and Investment. (2011a). *Chapter 3: Age-sex structure of the population in Vietnam. Hanoi: Vietnam population and housing census 2009*. Retrieved from http://www.gso.gov.vn/Modules/Doc_Download.aspx?DocID=13286.

Ministry of Planning and Investment. (2011b). *Chapter 4: Household structure. Hanoi: Vietnam population and housing census 2009*. Retrieved from http://www.gso.gov.vn/Modules/Doc_Download.aspx?DocID=13287

Moreira, M. M.; Blyde, J. S.; Rodriguez, D. M.; Martincus, C. V. (2013). *Peru: Road infrastructure and and regional exports with a challenging geography*. Retrieved from http://toofartoexport.com/peru.pdf

Moroney, J., Pezard, S., Miller, L., Engstrom, J., & Doll, A. (2013). *Lessons from Department of Defense disaster relief efforts in the Asia-Pacific region*. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_reports/RR100/ RR146/RAND_RR146.pdf

Moscow Metro. (2017). Metro at a glance. Retrieved from http://mosmetro.ru/press/ metropoliten-v-tsifrakh/index.php?sphrase_id=27068

Mosgorstat. (2010). 2010 National population census: Results. Retrieved from http://moscow.gks.ru/wps/wcm/connect/rosstat_ts/moscow/ru/ census_and_researching/census/national_census_2010/score_2010/ score_2010_default

Moss, M., & Townsend, A. (2005). *Telecommunications infrastructure in disasters: Preparing cities for crisis communications*. (Research Paper), Retrieved from https://www.nyu.edu/ccpr/pubs/NYU-DisasterCommunications1-Final.pdf

Mosvodokanal. (2017). Water sources. Retrieved from http://www.mosvodokanal.ru/ watersupply/sources/

MRT Jakarta. (2017). About project. Retrieved from http://jakartamrt.co.id/mengenai-proyek/

Mumbai Port Trust. (2017). *Yearwise, commodity wise traffic handled*. Retrieved from http://www.mumbaiport.gov.in/writereaddata/nmainlinkFile/File957.pdf

MWRA. (2017). Water use and system demand. Retrieved from http://www.mwra.state.ma.us/monthly/wsupdat/archivedemand.htm

Nairobi City Water and Sewrage Company. (2017). Sources. Retrieved from https://www.nairobiwater.co.ke/index.php/en/

Nairobi Government. (2017a). Emergency services. Retrieved from http://www.nairobi.go.ke/emergency-services/

Nairobi Government. (2017b). Explore Nairobi. Retrieved from http://www.nairobi.go.ke/home/explore-nairobi/

Nam, V. (2016, Feb. 7). HCMC: Per capita income increased by more than 70% in six years. Retrieved from http://www.thesaigontimes.vn/148424/TPHCM-Thu-nhap-binh-quan-dau-nguoi-tang-hon-70-trong-6-nam.html

Nasseri, L. (2014, Sept. 18). Iran may import water from Tajikistan to avert crisis. Retrieved from https://www.bloomberg.com/news/articles/2014-09-18/iran-may-import-water-from-tajikistan-to-avert-crisis

National Bureau of Statistics of China. (2015). National data. Retrieved from http://data.stats.gov.cn/english/easyquery.htm?cn=E0103

National Institute of Population Studies. (2013). *Pakistan demographic and health survey*. Retrieved from http://dhsprogram.com/pubs/pdf/FR290/FR290.pdf

National Institute of Statistics and Information. (2015). Statistics. Retrieved http://www.inei.gob.pe

National Water Agency. (2010a). Urban agglomerations. Retrieved from http://atlas.ana.gov.br/atlas/forms/AglomeradosUrbanos.aspx

National Water Agency. (2010b). Metropolitan region of Rio de Janeiro. Retrieved from http://atlas.ana.gov.br/atlas/forms/analise/RegiaoMetropolitana.aspx?rme=18

New Hampshire Hospital Association. (2017). Map of hospitals. Retrieved from http://www.nhha.org/index.php/nh-hospitals/map-of-hospitals

New York City Fire Department. (2014). *FDNY vital statistics*. Retrieved from http://www1.nyc.gov/assets/fdny/downloads/pdf/vital_stats_2014_cy.pdf

New York City Government. (2013). *Infrastructure*. Retrieved from http://www.nyc.gov/html/dot/downloads/pdf/2013-dot-sustainable-streets-5-infrastructure.pdf

New York City Government. (2015). *New York City drinking water supply and quality report*. Retrieved from http://www.nyc.gov/html/dep/pdf/wsstate15.pdf

New York City Government. (2017). NYPD precincts. Retrieved from http://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page

New York State. (2017a). Hospitals by region/county. Retrieved from https://profiles.health.ny.gov/hospital/county_or_region/region:hudson+valley

New York State. (2017b). Long Island aquifers. Retrieved from http://www.dec.ny.gov/lands/36183.html

Newark Fire Department. (2017). Fire stations. Retrieved from http://nfd.newarkpublicsafety.org/index.php

Newark Police Department. (2017). Precincts & divisions. Retrieved from http://npd.newarkpublicsafety.org/index.php

News Reports. (2013). Poverty in Karachi. Retrieved from http://www.borgenmagazine.com/poverty-in-karachi/

Nigerian Ports Authority. (2015). *Full year report*. Retrieved from https://www.nigerianports.org/dynamicdata/uploads/YearlyReports/2015-FULL-YEAR-REPORT.pdf

Nikolova, M. (2016, May 2). Two solutions to the challenges of population aging. Retrieved from https://www.brookings.edu/blog/up-front/2016/05/02/two-solutions-to-the-challenges-of-population-aging/

Nomis. (2011). Office of National Statistics Census 2011. Retrieved from https://www.nomisweb.co.uk/census/2011/data_finder

Open Data for Africa. (2017a). Nigeria data portal. Retrieved from http://nigeria.opendataforafrica.org/apps/atlas/Lagos

Open Data for Africa. (2017b). Egypt data portal. Retrieved from http://egypt.opendataforafrica.org/apps/atlas/Cairo

Open Data for Africa. (2017c). South Africa data portal. Retrieved from http://southafrica.opendataforafrica.org/apps/atlas/gauteng

Open Data for Africa. (2017d). Kenya data portal. Retrieved from http://kenya.opendataforafrica.org/apps/atlas/nairobi

Open Data for Africa. (2017e). Democratic Republic of the Congo data portal. Retrieved May 21, 2017, from http://drcongo.opendataforafrica.org/apps/atlas/Kinshasa

Orange County Water District. (2016). 2014–2015 Engineer's report on the groundwater conditions, water supply and basin utilization in the Orange County water district. Retrieved from https://www.ocwd.com/media/4260/2014-15-engineers-report.pdf

Orange County. (2017). Commands and divisions. Retrieved from http://www.ocsd.org/divisions/

PA Department of Health. (2013). Data from the annual hospital questionnaire. Retrieved from http://www.statistics.health.pa.gov/HealthStatistics/HealthFacilities/HospitalReports/Documents/Hospital_Report_2014_2015.pdf

Patna, B. (2013, May 11). India's demographic challenge wasting time. Retrieved from http://www.economist.com/news/briefing/21577373-india-will-soon-have-fifth-worlds-working-age-population-it-urgently-needs-provide

People's Republic of China. (2009). *Chongqing water supply project (CXXI-P121).* Retrieved from https://www2.jica.go.jp/en/evaluation/pdf/2008_CXXI-P121_4.pdf

Peru National Port Authority. (2016). *Movement of cargo in public and private terminals*. Retrieved from https://www.apn.gob.pe/site/wp-content/uploads/2017/02/pdf/7LNRXWVBZW9JXET8KKGTACYYUIURSDL2OCIF.pdf

Pew Research Center. (2017). Religious landscape study. Retrieved from http://www.pewforum.org/religious-landscape-study/

Philadelphia Police. (2017). Districts & units. Retrieved from https://www.phillypolice.com/districts-units/index.html

Philippine Government. (2016). 2016 Power statistics. Retrieved from https://www.doe.gov.ph/sites/default/files/pdf/energy_statistics/bgross_power_generation_by_plant_2016.pdf

Philippines Ports Authority. (2016). *Annual report: Summary on port statistics*. Retrieved from http://www.ppa.com.ph/?q=content/statistics-1

Port of London Authority. (2014). *Economic impact report summary*. Retrieved from http://www.pla.co.uk/assets/economicreport.pdf

Port of London Authority. (2017). *Port of London Authority handbook 2017*. Retrieved from http://www.pla.co.uk/assets/plahbook.pdf

Power Grid Company of Bangladesh LTD. (2016). Zonewise supply against highest generation(MW). Retrieved from https://pgcb.org.bd/PGCB/?a=pages/maxgen_display.php

Power Plants Around the World. (2016). Hydroelectric plants in Afghanistan. Retrieved from http://www.industcards.com/hydro-afghanistan.htm

Probe International Beijing Group. (2008). *Beijing's water crisis 1949–2008 Olympics*. Retrieved from http://www.chinaheritagequarterly.org/016/_docs/BeijingWaterCrisis1949-2008.pdf

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3–900051-07-0, Retrieved from http://www.R-project.org

Railway Technology. (2017). Bangkok metro, Thailand. Retrieved from http://www.railway-technology.com/projects/bangkok-metro/

Railway Technology. (2017). Lima Peru metro. Retrieved from http://www.railway-technology.com/projects/lima-metro

Reaves, J. (2010, Jan. 14). Fighting illiteracy in Chicago, with enthusiasm.*The New York Times*. Retrieved from http://www.nytimes.com/2010/01/15/education/15cncbooks.html?_r=0

Republic of the Philippines. (2017). Metro Manila water system. Retrieved from http://mwss.gov.ph/learn/metro-manila-water-supply-system/

Rio Port Aurhority. (2014). *Container movement 2008–2014*. Retrieved from http://www.portosrio.gov.br/downloads/files/estatistica/hist%C3%B3rico_de_cont%C3%AAineres_2008_a_2014_por_terminal_-_porto_do_rio.pdf

Rio Port Authority. (2016). *Superintendency of market planning and statistical intelligence management*. Retrieved from http://www.portosrio.gov.br/downloads/files/estatistica/mapa_resumo_dezembro_2016_-_porto_do_rio_de_janeiro.pdf

Rodrigues, M., & Silveira, D. (2016, Jul. 30). With Temer and Pezão, Line 4 of the subway in Rio is inaugurated. Retrieved from http://g1.globo.com/rio-de-janeiro/noticia/2016/07/temer-participa-de-inauguracao-da-linha-4-do-metro-no-rio.html

Salary Explorer. (2017). Salary survey Cairo. Retrieved from http://www.salaryexplorer.com/salary-survey.php?loc=730&loctype=3

San Francisco Public Utilities Commission. (2016a). 2015 urban water management plan. Retrieved from http://www.sfwater.org/modules/showdocument.aspx?documentid=9300

San Francisco Public Utilities Commission. (2016b). Annual report. Retrieved 19 2017 May, from Water Resources Division: https://www.joomag.com/magazine/water-resources-division-annual-report-fy-2015-2016/0871482001480967586?short

Sapol, S. (2016). *Megacity classification framework*. U.S. Army White Paper. FA49Q 16–002. Retrived from, Dr. Jeffrey Appleget, Senior Lecturer, Naval Postgraduate School.

Secretary of Energy and Minerals. (2015). 15 Largest consumer municipalities. Retrieved from http://dadosenergeticos.energia.sp.gov.br/PortalCEv2/Municipios/Eletricidade/M_Eletricidade.asp

Skuse, A. (n.d.). Ho Chi Minh City Vietnam's next door. Retrieved from http://chinainvestin.com/index.php/en/invest-in/spotlight/274/ho-chi-minh-city-vietnams-next-door

Smith, R. (2016, Dec. 30). Cyber attacks raise alarm for U.S. power grid. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/cyberattacks-raise-alarms-for-u-s-power-grid-1483120708

Sohel, K. (2014, Aug. 28). Rangpur has the highest poverty rate. *The Dhaka Tribune*. Retrieved from http://archive.dhakatribune.com/bangladesh/2014/aug/28/rangpur-has-highest-poverty-rate

Stafford County Sheriff. (2017). Stafford County Sheriff’s Office. Retrieved from http://www.staffordsheriff.com/

State of New Jersey. (2017). Find a hospital. Retrieved from http://www.nj.gov/health/healthfacilities/findhospital.shtml

Statistical Documentation and Service Center. (2017). Statistical data. Retrieved from http://www.gso.gov.vn/Default_en.aspx?tabid=766

Statistics Canada. (2011). Census profile. Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CSD&Code1=3520005&Geo2=PR&Code2=35&Data=Count&SearchText=Toronto&SearchType=Begins&SearchPR=01&B1=All&GeoLevel=PR&GeoCode=3520005

TCDD. (2015). Turkish state railways statistics: Mainline track lengths by provinces. Retrieved from http://www.tcdd.gov.tr/files/istatistik/20112015yillik.pdf

Tehran. (2017). Tehran traffic and transportation. Retrieved from http://en.tehran.ir/Default.aspx?tabid=103

Tehran Times. (2016). Iran’s power generation capacity reaches 75,365MW. Retrieved May 22, 2017, from http://www.tehrantimes.com/news/405353/Iran-s-power-generation-capacity-reaches-75-365-MW

TekCarta. (2017). Average household size (68 countries). Retrieved from https://www.nakono.com/tekcarta/databank/households-average-household-size/

Telangana Government. (2017). About HMR. Retrieved from http://hmrl.telangana.gov.in/about-hmr.html

Telecom Regulation Authority of India. (2017). *Telecom subscription reports*. Retrieved from http://www.trai.gov.in/sites/default/files/Press_Release_34_Eng_28_04_2017.pdf

Texas Department of Health and Human Services. (2003). Estimated Texas population, by area 2003. Retrieved from https://www.dshs.texas.gov/chs/popdat/ST2003.shtm

Texas DHHS. (2012). Utilization data for Texas acute care hospitals, by county 2012. Retrieved from https://www.dshs.texas.gov/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=8589984639

Texas Water Development Board. (2014). *Water use survey historical summary estimates, by county*. Retrieved from http://www2.twdb.texas.gov/ReportServerExt/Pages/ReportViewer.aspx?%2fWU%2fSumFinal_CountyReport&rs:Command=Render

Texas Water Development Board. (2017). Major aquifers & minor aquifers. Retrieved from http://www.twdb.texas.gov/groundwater/aquifer/major.asp

TomTom International. (2016). Rio de Janeiro. Retrieved from https://www.tomtom.com/en_gb/trafficindex/city/rio-de-janeiro

Toronto. (2012). 2012 Update to the road classification system. Retrieved from http://www1.toronto.ca/wps/portal/contentonly? vgnextoid=d827a84c9f6e1410VgnVCM10000071d60f89RCRD&vgnextchannel=6f2c4074781e1410VgnVCM10000071d60f89RCRD

Toronto. (2013). Poverty, housing and homelessness. Retrieved from https://www1.toronto.ca/City%20Of%20Toronto/Affordable%20Housing%20Office/Shared%20Content/pdf/poverty-factsheet.pdf

Toronto. (2016). *Economic overview 2016*. Retrieved from http://www1.toronto.ca/City%20Of%20Toronto/Economic%20Development%20&%20Culture/Business%20Pages/Reports%20&%20Data%20Centre/Economic%20Overview%20-%20EDC%20Format%20-%20Nov%208,%202016_ecdevdata.pdf

Toronto. (2017a). Map of Toronto: Community hospitals. Retrieved from http://map.toronto.ca/maps/map.jsp?app=TorontoMaps_v2

Toronto. (2017b). Toronto facts: Infrastructure. Retrieved from http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=91d5f937de453410VgnVCM10000071d60f89RCRD&vgnextchannel=57a12cc817453410VgnVCM10000071d60f89RCRD

Toronto. (2017c). Toronto water supply. Retrieved from http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=46d807ceb6f8e310VgnVCM10000071d60f89RCRD&vgnextchannel=b8011b9c6a85f310VgnVCM10000071d60f89RCRD

Toronto Police. (2017). Toronto police service telephone directory. Retrieved from http://www.torontopolice.on.ca/directory/

Toronto Transportation Commission. (2013). 2013 TTC operating statistics. Retrieved from http://www.ttc.ca/About_the_TTC/Operating_Statistics/2013.jsp

Tortajada, C., Variz, O., Lundqvist, J., & Biswas, A. (2006). Water management for large cities. New York, NY: Rutledge Taylor & Francis Group.

Tour Egypt. (2017). Cairo, Egypt statistics. Retrieved May 21, 2017, from http://www.touregypt.net/cairo/cairostatistics.htm

Trust for London and New Policy Institute. (2015). Poverty before and after housing costs. Retrieved from http://www.londonspovertyprofile.org.uk/indicators/topics/income-poverty/poverty-before-and-after-housing-costs/

Tuoitre News. (2014, Jan. 16). Ho Chi Minh City raises poverty line by 33%. Retrieved May 22, 2017, from http://tuoitrenews.vn/society/16940/ho-chi-minh-city-raises-poverty-line-by-33

Turkey Statistical Institute. (2013). Population and housing census 2011. Retrieved from http://www.turkstat.gov.tr/PreHaberBultenleri.do?id=15843

Turkey Statistical Institute. (2017a). Statistics by subject. Retrieved from http://www.tuik.gov.tr/UstMenu.do?metod=kategorist

Turkey Statistical Institute. (2017b). Municpal water statistics. Retrieved from https://biruni.tuik.gov.tr/medas/?kn=121&locale=tr

Tutuko, K. (1998). Jakarta water supply. Retrieved from https://www.pecc.org/resources/infrastructure-1/1227-jakarta-water-supply-how-to-implement-a-sustainable-process-1/file

UK Health Centre. (2017). Hospitals in London and UK. Retrieved from http://www.healthcentre.org.uk/hospitals/find.html

UN Data. (2017). Democratic Republic of the Congo. Retrieved from http://data.un.org/CountryProfile.aspx?crName=democratic%20republic%20of%20the%20congo

UNICEF. (2013). At a glance: Mexico. Retrieved from https://www.unicef.org/infobycountry/mexico_statistics.html

United Kingdom. (2011). Religion. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/religion/articles/religioninenglandandwales2011/2012-12-11

United Kingdom. (2015). Statistics. from https://www.gov.uk/government/statistics

United Kingdom. (2017). Office for National Statistics. Retrieved from https://www.ons.gov.uk/

United Nations, IMO. (2017). International Maritime Organization. Retrieved from https://business.un.org/en/entities/13

United Nations. (2015). World Population Prospects: 2015 Revision. Retrieved from https://esa.un.org/unpd/wpp/publications/files/key_findings_wpp_2015.pdf

United States Army. (2013). *Operational environment considerations for training and education development*. Retrieved fromhttp://www.benning.army.mil/mssp/security%20topics/Potential%20Adversaries/content/pdf/tc7-102.pdf

United States Army. (2014). *Megacities and the United States Army*. Strategic Studies Group. Retrieved from https://www.army.mil/e2/c/downloads/351235.pdf

United States Census Bureau. (2015a). Quick facts. Retrieved from https://www.census.gov/quickfacts/

United States Census Bureau. (2015b). Presence and type of internet subscriptions in household Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t#none

United States Department of Commerce. (2015). GDP & personal income, current dollars. Retrieved from https://www.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=3#reqid=70&step=1&isuri=1

United States Department of Transportation. (2012). Public road length, miles by ownership: 2011. Retrieved from https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/state_transportation_statistics/state_transportation_statistics_2012/html/table_01_02.html

United States Department of Transportation. (2015). U.S. waterborne foreign trade by U.S. Customs Ports 2000–2015. Retrieved from https://www.marad.dot.gov/resources/data-statistics/

United States Department of Transportation. (2017). Airports. Retrieved, from https://www.faa.gov/airports/airport_safety/airportdata_5010/menu/

United States Energy Information Administration. (2017). State profiles and estimates. Retrieved from https://www.eia.gov/state/

United States Fire Administration. (2017). Download national list. Retrieved from https://apps.usfa.fema.gov/registry-download/main/download

United States Marine Corps. (2014). Current operations brief. Retrieved from http://www.hqmc.marines.mil/Portals/138/HQMC%20Current%20Ops%20Brief%2023%20Oct%202014.pptx

United States Marine Corps. (2015). *Security environment forecast 2030–2045*. Futures Directorate, Combat Development & Integration. Retrieved from http://www.mcwl.marines.mil/Portals/34/Documents/2015%20MCSEF%20-%20Futures%202030-2045.pdf

University of British Columbia. (1999). Water sources in Shanghai. Retrieved from http://www.chs.ubc.ca/china/shanghai.pdf

University of Tokyo. (n.d.). Asian Metropolis: Dhaka. Retrieved May 22, 2017, from http://www.urban.t.u-tokyo.ac.jp/asianmetropolis/Dhaka.pdf

US Hospitals. (2017). Delaware hospitals. Retrieved from http://www.ushospital.info/Delaware.htm

USAID. (2003). Household population and housing characteristics. Retrieved from http://dhsprogram.com/pubs/pdf/FR151/02Chapter02.pdf

USAID. (2016). Kabul urban water supply KUWS) overview. Retrieved from https://www.usaid.gov/news-information/fact-sheets/kabul-urban-water-supply-kuws

USGS. (2009). Conceptual model of water resources in the Kabul basin, Afghanistan. Retrieved May from https://pubs.usgs.gov/sir/2009/5262/

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

Van Leeuwen, C. J., Nguyen, D. P., & Dieperink, C. (2015). The challenges of water governance in Ho Chi Minh City. *Integrated Environmental Assessment and Management, 9999(9999), 1–8*. Retrieved from https://www.watershare.eu/wp-content/uploads/10.1002_ieam.1664.pdf

Varinsky, D. (2016, May 20). Saudi Arabia is building the largest urban-transit system ever made from scratch. Retrieved from http://www.businessinsider.com/saudi-arabias-riyadh-metro-will-be-the-largest-urban-transit-system-ever-built-from-scratch-2016-5

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S: Fourth edition*. New York, New York: Springer Science+Business Media.

Vietnam Seaports Association. (2015a). Saigon New Port. Retrieved from http://www.vpa.org.vn/english/members/south/saigonnew.htm

Vietnam Seaports Association. (2015b). Saigon Port. Retrieved from http://www.vpa.org.vn/english/members/south/saigon.htm

Virginia American Water. (2017). Virginia water systems. Retrieved from https://amwater.com/vaaw/water-quality/virginia-water-systems/

Washington, DC. (2017a). Metropolitan police department. Retrieved from https://mpdc.dc.gov/

Washington, DC. (2017b). Hospital facility directory. Retrieved from http://doh.dc.gov/node/173192

Wisconsin Department of Health Services. (2017). Hospitals directory. Retrieved from https://www.dhs.wisconsin.gov/guide/hospital.htm

Wickham, H. (2009). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York, 2009. Retrived from http://ggplot2.org

Wombell, J. (2009). *Army support to the Hurricane Katrina disaster*. Fort Leavenworth: Combat Studies Institute Press. Retrieved from http://usacac.army.mil/cac2/cgsc/carl/download/csipubs/wombwell.pdf

World Bank. (2008). *Economic impacts of sanitation in Indonesia*. Retrieved from https://www.wsp.org/sites/wsp.org/files/publications/esi_indonesia.pdf

World Bank. (2015). *Water supply and sanitation in the Philippines*. Retrieved from https://www.wsp.org/sites/wsp.org/files/publications/WSP-Philippines-WSS-Turning-Finance-into-Service-for-the-Future.pdf

World Bank. (2016, Oct. 8). Bangladesh: Improving water supply and sanitation. Retrieved from http://www.worldbank.org/en/results/2016/10/07/bangladesh-improving-water-supply-and-sanitation

World Bank. (2017a). Air transport, freight. Retrieved from http://data.worldbank.org/indicator/IS.AIR.GOOD.MT.K1

World Bank. (2017b). Air transport, passengers carried. Retrieved from http://data.worldbank.org/indicator/IS.AIR.PSGR

World Bank. (2017c). Improved sanitation facilities. Retrieved from http://data.worldbank.org/indicator/SH.STA.ACSN?locations=US

World Port Source. (2017). Port of Istanbul. Retrieved from http://www.worldportsource.com/ports/commerce/TUR_Port_of_Istanbul_3090.php

WSSC. (2017). WSSC dams and reservoirs. Retrieved from https://www.wsscwater.com/dams

Yandex. (2016). Internet development in Russian Regions. Retrieved from https://yandex.ru/company/researches/2016/ya_Internet_regions_2016#polzovanieinternetomvraznyxgorodax

Zykov, T. (2015, Apr. 29). Poverty is not the threshold. Retrieved from https://rg.ru/2015/04/29/bednost.html

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California